Can Language Models be Biomedical Knowledge Bases?

Mujeen Sung[1] **Jinhyuk Lee**[1,2†] **Sean S. Yi**[1]
Minji Jeon[3] **Sungdong Kim**[4] **Jaewoo Kang**[1†]
Korea University[1] Princeton University[2]
Icahn School of Medicine at Mount Sinai[3] NAVER AI Lab[4]
{mujeensung,jinhyuk_lee,seanswyi,kangj}@korea.ac.kr
minji.jeon@mssm.edu
sungdong.kim@navercorp.com

Abstract

Pre-trained language models (LMs) have become ubiquitous in solving various natural language processing (NLP) tasks. There has been increasing interest in what knowledge these LMs contain and how we can extract that knowledge, treating LMs as knowledge bases (KBs). While there has been much work on probing LMs in the general domain, there has been little attention to whether these powerful LMs can be used as domain-specific KBs. To this end, we create the BIOLAMA benchmark, which is comprised of 49K biomedical factual knowledge triples for probing biomedical LMs. We find that biomedical LMs with recently proposed probing methods can achieve up to 18.51% Acc@5 on retrieving biomedical knowledge. Although this seems promising given the task difficulty, our detailed analyses reveal that most predictions are highly correlated with prompt templates without any subjects, hence producing similar results on each relation and hindering their capabilities to be used as domain-specific KBs. We hope that BIOLAMA can serve as a challenging benchmark for biomedical factual probing.[1]

1 Introduction

Recent success in natural language processing can be largely attributed to powerful pre-trained language models (LMs) that learn contextualized representations of words from large amounts of unstructured corpora (Peters et al., 2018; Devlin et al., 2019). There have been recent works in probing how much knowledge these LMs contain in their parameters (Petroni et al., 2019) and how to effectively extract such knowledge. (Shin et al., 2020; Jiang et al., 2020b; Zhong et al., 2021).

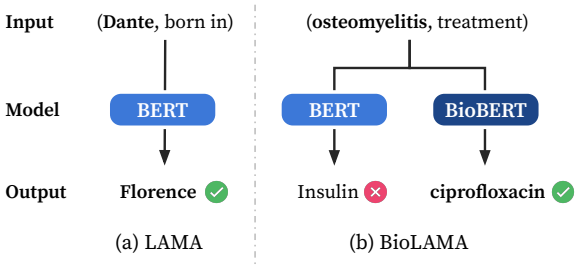

Figure 1: Comparison of LAMA (Petroni et al., 2019) and BIOLAMA. (a) LAMA tests general knowledge of LMs. (b) BIOLAMA tests expert-level biomedical knowledge of LMs such as a treatment for a disease.

While factual probing of LMs has attracted much attention from researchers, a more practical application would be to leverage the power of domain-specific LMs (Beltagy et al., 2019; Lee et al., 2020) as domain knowledge bases (KBs). Unlike recent works that probe general domain knowledge, we ask whether it is also possible to retrieve expert knowledge from LMs. Specifically, we tune our focus on factual knowledge probing for the biomedical domain as shown in Figure 1.

To inspect the potential utility of LMs as biomedical KBs, we create and release the **Bio**medical **LA**nguage **M**odel **A**nalysis (BIOLAMA) probe. BIOLAMA consists of 49K biomedical factual triples whose relations have been manually curated from three different knowledge sources: the Comparative Toxicogenomics Database (CTD), the Unified Medical Language System (UMLS), and Wikidata. While our biomedical factual triples are inherently more difficult to probe (see Table 1 for examples), BIOLAMA also poses technical challenges such as multi-token object decoding.

Initial probing results on BIOLAMA show that the best performing LM achieves up to 7.28% Acc@1 and 18.51% Acc@5, and outperforms an information extraction (IE) baseline (Lee et al., 2016). Although this result seems promising, we find that their output distributions are largely biased to a

[1] https://github.com/dmis-lab/BioLAMA
[†] Corresponding authors.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4723–4734
November 7–11, 2021.

	Relation Name	Manual Prompt	Object Answer
LAMA	**# Relations**: 41	**# Entity Types**: 25* **# Triples**: 41k	**Sources**: Wikidata
	place of birth	Dante was born in [Y].	Florence
	place of death	Adolphe Adam died in [Y].	Paris
	official language	The official language of Mauritius is [Y].	English
BIOLAMA	**# Relations**: 36	**# Entity Types**: 12 **# Triples**: 49K	**Sources**: CTD, UMLS, Wikidata
	medical condition treated	Amantadine has effects on [Y].	Parkinson's disease, ...
	symptoms	Hepatitis has symptoms such as [Y].	abdominal pain, ...
	affects binding	Nicotine binds to [Y].	CHRNA4, CHRNB2, ...

Table 1: Comparison of LAMA (T-REx) and BIOLAMA with their statistics. For each dataset, we also show the examples of relations and their corresponding manual prompts and answers. The underlined entities are subjects and [Y] refers to the object to be predicted. *: obtained from Cao et al. (2021).

small number of entities in each relation. Along this line, we use two metrics, prompt bias (Cao et al., 2021) and synonym variance, to investigate the behavior of LMs as KBs. Our analysis shows that while LMs seem to be more aware of synonyms than the IE baseline, they output highly biased predictions given the prompt template of each relation. Our result calls for better LMs and probing methods that can retrieve rich but still useful biomedical entities.

2 BIOLAMA

In this section, we detail the construction of BIOLAMA including the data curation process and pre-processing steps. Statistics and examples of BIOLAMA are shown in Table 1 along with those from LAMA (Petroni et al., 2019).

2.1 Knowledge Sources

CTD The CTD[2] is a public biomedical database on relationships and interactions between biomedical entities such as diseases, chemicals, and genes (Davis et al., 2020). It provides both manually curated and automatically inferred triples in English, and we only use the manually curated triples for a better quality of our dataset. We use the April 1st, 2021 version of the CTD.

UMLS The UMLS Metathesaurus[3] is a large-scale database that provides information regarding various concepts and vocabularies in the biomedical domain (Bodenreider, 2004). We use the 2020AB version of the UMLS. The UMLS provides entity names in various languages and we use the ones in English.

Dataset	Obj in Sbj (%)	# Object Subwords
LAMA	12.81	1.00
LAMA-UHN	**0.00**	1.00
X-FACTR	6.35	3.07
BIOLAMA	**0.00**	**4.52**

Table 2: Comparison of probing benchmarks: ratio of subjects with objects as substrings, and the average subword numbers of object entities. We compare these two aspects of BIOLAMA to LAMA, LAMA-UHN (Poerner et al., 2020) and X-FACTR (Jiang et al., 2020a).

Wikidata Wikidata[4] is a public KB with items across various domains. Following the previous works (Turki et al., 2019; Waagmeester et al., 2020), we retrieve biomedical entities and relations using SPARQL queries. We use the dump of the January 25th, 2021 version. Similar to the UMLS, we use entity names in English.

2.2 Data Pre-processing

From our initial factual triples from the knowledge sources above, we apply several pre-processing steps to further improve the quality of BIOLAMA. First, considering the trade-off between the coverage and difficulty of probing, we restrict the lengths of entities to be $\leq$10 subwords, which covers 90% of the entities.[5] Note that LAMA only contains single-token objects, which makes the task easier, but less practical. Following Poerner et al. (2020), we also discard easy triples where objects are substrings of the paired subjects (e.g., "iron deficiency"-"iron"), which prevents trivial solutions using the surface forms of the subjects. For each relation, we split samples into training, development, and test sets with a 40:10:50 ratio. The training set is provided for learning or finding good

[2] http://ctdbase.org/
[3] https://www.nlm.nih.gov/research/umls/
[4] https://wikidata.org
[5] Based on the `BERT-base-cased` tokenizer.

Source	IE	BERT		BioBERT		Bio-LM	
		Manual	Opti.	Manual	Opti.	Maual	Opti.
CTD (11.13%)	**5.06 / 12.15**	0.06 / 1.20	3.56 / 6.97	0.42 / 3.25	4.82 / 9.74	1.77 / 7.30	2.99 / 10.19
UMLS (9.67%)	3.53 / 6.99	0.82 / 1.99	1.44 / 3.65	1.16 / 3.82	5.08 / 13.28	3.44 / 8.88	**8.25 / 20.19**
Wikidata (5.76%)	7.03 / 15.55	1.16 / 6.04	3.29 / 8.13	3.67 / 11.20	4.21 / 12.91	**11.97 / 25.92**	10.60 / 25.15
Average	5.21 / 11.56	0.86 / 3.08	2.76 / 6.25	1.75 / 6.09	4.70 / 11.98	5.72 / 14.03	**7.28 / 18.51**

Table 3: Main experimental results on BIOLAMA. We report Acc@1/Acc@5 of each model including the macro average across three different knowledge sources. We also report ratios of the majority objects in each knowledge source (averaged over its relations) in the parentheses. Highest and second-highest scores are **boldfaced** and underlined, respectively. Manual: manual prompt. Opti.: OptiPrompt. The results of OptiPrompt are the mean of 5 runs with different seeds. See Appendix E for the performance on each relation.

prompts for each relation. More details on pre-processing steps are available in Appendix A.

After the pre-processing, we are able to obtain 22K triples with 15 relations from the CTD, 21.2K triples with 16 relations from the UMLS, and 5.8K triples with 5 relations from Wikidata (see Appendix B for the detailed statistics). In Table 2, we compare various probing benchmarks with BIOLAMA. By design, BIOLAMA has no objects that are substrings of their subjects and object entities are much longer on average, which makes our benchmark challenging but much more practical.

Evaluation Metric We use top-k accuracy (Acc@k), which is 1 if any of the top k object entities are included in the annotated object list, and is 0 otherwise. We use both Acc@1 and Acc@5 since most biomedical entities are related to multiple biomedical entities (i.e., *N-to-M* relations).

3 Experiment

3.1 Models

Information Extraction Many biomedical NLP tools rely on automated IE systems that can provide relevant entities or articles given a query. In this work, we use the Biomedical Entity Search Tool (BEST) (Lee et al., 2016)[6] as an IE system and compare it with LM-based probing methods. BEST incorporates biomedical entities when building their search index over PubMed, a large-scale biomedical corpus, and returns biomedical entities given a keyword-based query. To fully make use of BEST, we create AND queries using a subject entity and a lemmatized relation name (e.g., "(meclozine) AND (medical condition treat)"), and use retrieved entities as its predictions.

Language Models We use one general-domain LM and two biomedical LMs: BERT (Devlin et al., 2019), BioBERT (Lee et al., 2020),[7] and Bio-LM (Lewis et al., 2020).[8] BioBERT and Bio-LM are both pre-trained over PubMed. While Bio-LM also uses a custom vocabulary learned from PubMed, BioBERT uses the same vocabulary as BERT, which enables the continual learning of BioBERT initialized from BERT.

3.2 Probing Methods

Prompts We use a fill-in-the-blank cloze statement (i.e., a "prompt") for probing and choose two different methods of prompt generation: manual prompts (Petroni et al., 2019) and OptiPrompt (Zhong et al., 2021). For each relation, we first create manual prompts with domain experts (Appendix C). On the other hand, OptiPrompt automatically learns continuous embeddings that can better extract factual knowledge for each relation, which are trained with our training examples. Following Zhong et al. (2021), we initialize the continuous embeddings with the embeddings of manual prompts, which worked consistently better than random initialization in our experiments.

Multi-token Object Decoding Since the majority of entities in BIOLAMA are made up of multiple tokens, we implement a multi-token decoding strategy following Jiang et al. (2020a). Among their decoding methods, we use the confidence-based method which produced the best results. The confidence-based method greedily decodes output tokens sorted by the maximum logit in each token position. Note that we do not restrict our output spaces by any pre-defined sets of biomedical entities since we are more interested in how accurately the LMs contain biomedical knowledge in

[6] https://best.korea.ac.kr/

[7] Since existing checkpoints of BioBERT do not contain LM heads for probing, we pre-train another BioBERT (`biobert-base-cased-v1.2`), which is the same as the previous version of BioBERT but with an LM head.

[8] `RoBERTa-base-PM-Voc`

Relation ID	Subject	Top 5 Predictions
UMLS - UR254 (27.71 / 38.41)	[X] has symptoms such as [Y].	
	Pituicytoma	**headache**, headaches, pain, bone pain, pain and bleeding
	Intravascular fasciitis	**pain**, pain and swelling, swelling and pain, swelling, edema
	Microfollicular adenoma	headache, epistaxis, pruritus, itching, flushing
	Parosteal Osteosarcoma	pain, bone pain, pain and swelling, swelling and pain, pain and bleeding
CTD - CG4 (8.42 / 20.59)	[X] results in increased activity of [Y] protein.	
	Dieldrin	**ESR1**, **NR1I2**, NR3C1, CASP1, PPARÎ³
	isofenphos	**ESR1**, NR1I2, PPARÎ³, CYP1A2, CDKN1A1
	Dithiothreitol	ESR1, NR1I2, NR3C1, NR1I1, CASP1
	Indigo Carmine	ESR1, NR1I2, CYP1A2, NR3C1, CASP1
Wikidata - P2176 (20.14 / 39.57)	The standard treatment for patients with [X] is a drug such as [Y].	
	Haverhill fever	**doxycycline**, ciprofloxacin, penicillin, erythromycin, azithromycin
	influenza	**zanamivir**, interferon, **peramivir**, oseltamivir or peramivir, doxycycline
	cryptosporidiosis	amphotericin B, praziquantel, itraconazole, albendazole, fluconazole
	tremor	pilocarpine, baclofen, botulinum toxin, diazepam, clonazepam

Table 4: Top 5 predictions of Bio-LM (w/ OptiPrompt) given each prompt and different subjects. For each relation, we also report its Acc@1/Acc@5. Correct predictions are in boldface. For more examples, see Appendix F.

an unconstrained setting.[9] See Appendix D for the implementation details of our decoding method.

3.3 Main Results

Experimental results on BIOLAMA are summarized in Table 3. First, BioBERT and Bio-LM are both able to retrieve factual information better than BERT, which demonstrates the effectiveness of domain-specific pre-training. Also, Bio-LM shows consistently better performance than BioBERT (BERT < BioBERT < Bio-LM). We believe that this may be attributed to the custom vocabulary of Bio-LM learned from a biomedical corpus. Using OptiPrompt also shows consistent improvement over manual prompts in all LMs. Notably, the IE system is able to achieve the best performance on the CTD relations, but performs worse than BioBERT and Bio-LM on the UMLS and Wikidata relations.

While we are able to achieve 18.51% Acc@5 with Bio-LM (w/ OptiPrompt) on average, note that the average Acc@1s on the CTD and UMLS relations are lower than majority voting (e.g., 9.67% (majority) vs. 8.25% Acc@1 (Bio-LM) in UMLS), which shows the difficulty of accurately extracting biomedical facts from these models.

4 LMs are Not Biomedical KBs, Yet

In this section, we thoroughly inspect the predictions of Bio-LM (w/ OptiPrompt) and quantitatively characterize the behavior of each model. Our analyses suggest that we might need stronger biomedical LMs and probing methods to make use of these LMs as domain-specific knowledge bases.

4.1 Predictions

In Table 4, we present two correct and two incorrect predictions for three different relations where Bio-LM (w/ OptiPrompt) achieves high accuracy. One aspect that stands out is that predictions tend to be highly biased towards a few objects (e.g., "headache", "pain", or "ESR1"). Motivated by this observation, we further measure two metrics that can characterize the behavior of each model in detail: prompt bias and synonym variance.

4.2 How Biomedical LMs Predict

Prompt Bias To serve as accurate KBs, LMs must make appropriate object entity predictions given the input subject entity. Cao et al. (2021) quantified prompt biases by measuring how insensitive LMs are to input subjects. For each relation, we first obtain the probability histogram of each unique object entity being a top-1 prediction *when the subject is given*. For example, if one relation has 100 test samples and "pain" appears 20 times as its top-1 prediction, the probability mass of "pain" becomes 20%. At the same time, we calculate the probability distribution over unique object entities *when the subject is masked out* (see Figure 2). For instance, a model might assign 30% to "pain" even when the subject is masked out from the prompt. Prompt bias is the Pearson's correlation coefficient

[9] Using a pre-defined set of object entities removes the necessity of using complicated decoding strategies and will possibly improve the probing accuracy as well, which we leave as future work.

Original Input

The treatment for **arthritis** is [Y].

Input for Prompt Bias

The treatment for **[MASK]** is [Y].

Input for Synonym Variance

The treatment for **joint inflammation** is [Y].
A synonym of 'arthritis'

Figure 2: Examples of inputs for measuring prompt bias and synonym variance. We use a `[MASK]` token for the subject when measuring prompt bias, and replace each subject into their synonyms when measuring synonym variance.

between these two distributions, which indicates how biased the model is to a prompt. A lower prompt bias means that a model is giving less biased predictions for each relation (i.e., prompt).

Synonym Variance Biomedical entities often have a number of synonyms, which are often leveraged for modeling biomedical entity representations (Sung et al., 2020). Hence, it is important that predictions over our factual triples do not change when the input subject is replaced by its synonyms. To assess this aspect, we propose a metric called synonym variance, which measures how much each prediction changes when the subject is replaced with its synonyms (see Figure 2). We create 10 copies of our datasets by replacing the subjects with one of their synonyms chosen randomly. Synonym variance is the standard deviation of Acc@5 calculated from these new test sets. Lower synonym variance means that a model is giving more consistent predictions even with different synonyms.

Results Figure 3 shows the results of prompt biases in four different models. Compared to the IE system, the LMs have relatively higher correlations (over 0.6) meaning that their predictions are more biased towards the prompts. On the other hand, in Figure 4, LMs show relatively lower standard deviations over variations of synonyms than the IE system does. While this can be interpreted that the LMs are more robust to synonym variations, it might also be the result of strong biases in LMs on their prompts. For example, while BERT has the smallest synonym variance, it has the largest prompt bias, meaning that it is not a synonym-aware model, but just a highly biased model.

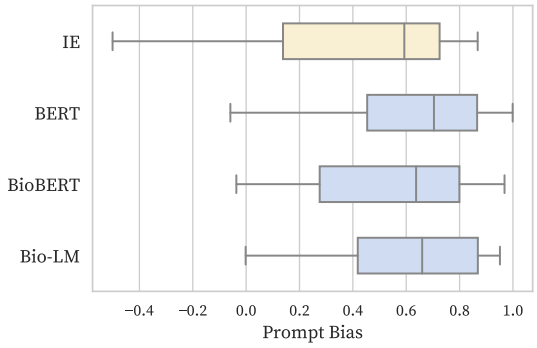

Figure 3: Prompt bias of each model. Low prompt bias means that a model is less biased on each prompt. See §4.2 for more details of the metric.

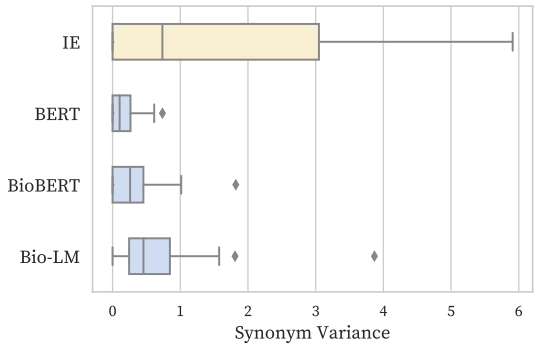

Figure 4: Synonym variance of each model. Low synonym variance means that a model gives consistent predictions when the subjects are changed to synonyms. See §4.2 for more details of the metric.

5 Conclusion

In this work, we explore the possibility of using LMs as biomedical KBs. To this end, we release BIOLAMA as a probing benchmark to measure how much biomedical knowledge can be extracted from LMs. While biomedical LMs can extract useful facts to some extent, our analysis shows that this is largely due to their predictions being biased towards certain prompts. In future work, we plan to overcome the underlying challenges in BIOLAMA and improve the probing accuracy of LMs.

Acknowledgements

This work was supported in part by the ICT Creative Consilience program (IITP-2021-0-01819) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation), National Research Foundation of Korea (NRF-2020R1A2C3010638, NRF-2014M3C9A3063541), and Hyundai Motor Chung Mong-Koo Foundation. We thank the anonymous reviewers for their insightful comments.

Ethical Considerations

The aim of factual probing is to verify how much knowledge can be retrieved from language models pre-trained using large amoung of corpora. Due to a lack of data for factual probing in the biomedical domain, we collected data from widely used knowledge sources: the CTD, the UMLS, and Wikidata. Although these data have undergone inspection by domain experts, biomedical knowledge is continuously growing and therefore we cannot guarantee that this biomedical knowledge is absolute. Furthermore, without careful inspection, outputs of these LMs should not be considered as a means of drug recommendation or any other medical activity. We caution future researchers when using BIOLAMA to keep this caveat in mind.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1860–1874, Online. Association for Computational Linguistics.

Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. 2020. Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Research*, 49(D1):D1138–D1143.

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. Association for Computational Linguistics.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020a. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020b. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

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*, 36(4):1234–1240.

Sunwon Lee, Donghyeon Kim, Kyubum Lee, Jaehoon Choi, Seongsoon Kim, Minji Jeon, Sangrak Lim, Donghee Choi, Sunkyu Kim, Aik-Choon Tan, et al. 2016. Best: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PloS one*, 11(10):e0164680.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-BERT: Efficient-yet-effective entity embeddings for BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 803–818, Online. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with

Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650, Online. Association for Computational Linguistics.

Houcemeddine Turki, Thomas Shafee, Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, Denny Vrandečić, Diptanshu Das, and Helmi Hamdi. 2019. Wikidata: A large-scale collaborative ontological medical database. *Journal of biomedical informatics*, 99:103292.

Andra Waagmeester, Gregory Stupp, Sebastian Burgstaller-Muehlbacher, Benjamin M Good, Malachi Griffith, Obi L Griffith, Kristina Hanspers, Henning Hermjakob, Toby S Hudson, Kevin Hybiske, et al. 2020. Science forum: Wikidata as a knowledge graph for the life sciences. *Elife*, 9:e52614.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

A Pre-processing of BIOLAMA

After applying basic pre-processing steps in §2, we aggregate samples with the same subject and relation, which makes each sample contain multiple object answers (e.g., subj="COVID-19", relation="symptoms of", obj={headache, cough, fever, ...}). We also set the maximun number of triples in each relation as 2,000 while removing relations having less than 500 triples, which are mostly less useful to extract (e.g., "affects methylation" in CTD) or too complicated (e.g., "positive therapeutic predictor" in Wikidata) according to our manual inspection with domain experts. For the UMLS, out of 974 relations, we select 16 relations that are considered to be the most important by domain experts. To mitigate the class imbalance problem in object entities, we also undersample highly frequent object entities to be as frequent as the fifth frequent object entity in each relation.

B Statistics of BIOLAMA

The CTD split has a total of 22,017 samples, the UMLS split a total of 21,164, and the Wikidata split a total of 5,855 samples. This sums up to a total of 49,036 samples. Table 5 displays the number of samples in each train/dev/test split of each relation.

C Manual Prompts

We create multiple manual prompts with the help of domain experts' insight on each relation in BIOLAMA and select the best performing prompts on the development set. Selected prompts for the relations are listed in Table 6.

D Implementation Details

For confidence-based decoding (Jiang et al., 2020a), we use the open-source code provided by the authors[10] and make slight changes for BIOLAMA. We set the beam size to 5 to get the top 5 predictions and the number of masks to 10. We also set the iteration method to "None" as additional iteration did not help to increase the performance.

For OptiPrompt (Zhong et al., 2021), we modify the open-source code provided by the authors[11] to allow training over the multi-token objects. We set the learning rate to 3e-3 and the mini-batch size to 16. We train OptiPrompt for 10 epochs and select the best checkpoint based on Acc@1 on the development set. It takes 3 hours to test all samples with manual prompts and 8 hours to train and test with OptiPrompt using 1 Titan X (12GB) GPU.

E Result on Each Relation

In addition to the averaged performances presented in Table 3, we present Acc@1 and Acc@5 on each relation in Table 7.

F More Prediction Examples

We provide more examples on 8 relations where Bio-LM (w/ OptiPrompt) achieves decent top-1 accuracy in Table 8.

[10] https://github.com/jzbjyb/X-FACTR
[11] https://github.com/princeton-nlp/OptiPrompt

Relation ID	Relation Name	Subject	Object	Train	Dev	Test
CTD						
CD1	therapeutic	chemical	disease	756	189	945
CD2	marker/mechanism	chemical	disease	723	181	905
CG1	decreases expression	chemical	protein	550	137	688
CG17	increases expression	chemical	mRNA	740	186	926
CG18	increases expression	chemical	protein	680	170	851
CG2	decreases activity	chemical	protein	718	179	898
CG21	increases phosphorylation	chemical	protein	206	51	258
CG4	increases activity	chemical	protein	541	135	677
CG6	decreases expression	chemical	mRNA	648	163	811
CG9	affects binding	chemical	protein	352	89	441
CP1	decreases	chemical	phenotype	504	127	631
CP2	increases	chemical	phenotype	591	148	739
CP3	affects	chemical	phenotype	360	90	451
GD1	marker/mechanism	gene	disease	728	182	911
GP1	association	gene	pathway	704	176	881
UMLS						
UR116	clinically associated with	disease	disease	668	167	835
UR124	may treat	disease	chemical	463	116	580
UR173	causative agent of	disease	vertebrate	512	128	640
UR180	is finding of disease	disease	body substance	346	87	434
UR211	biological process involves gene product	gene	function	650	162	813
UR214	cause of	disease	disease	459	115	574
UR221	gene mapped to disease	disease	gene	241	61	302
UR254	may be finding of disease	disease	symptom	672	169	841
UR256	may be molecular abnormality of disease	disease	genetic aberrant	244	62	306
UR44	may be prevented by	chemical	disease	452	113	566
UR45	may be treated by	chemical	disease	772	193	965
UR48	physiologic effect of	chemical	disease	700	176	876
UR49	mechanism of action of	chemical	function	615	154	769
UR50	therapeutic class of	chemical	type	663	166	829
UR588	process involves gene	gene	disease	621	156	777
UR625	disease has associated gene	gene	disease	381	96	477
Wikidata						
P2175	medical condition treated	chemical	disease	621	155	777
P2176	drug used for treatment	disease	chemical	448	112	561
P2293	Genetic association	gene	disease	678	170	849
P4044	therapeutic area	chemical	disease	304	76	380
P780	symptoms	disease	symptom	289	73	362
Total				19,600	4,910	24,526

Table 5: Detailed statistics of BioLAMA for each relation.

Relation ID	Relation Name	Manual Prompt
CTD		
CD1	therapeutic	[X] prevents diseases such as [Y].
CD2	marker/mechanism	[X] exposure is associated with significant increases in diseases such as [Y].
CG1	decreases expression	[X] treatment decreases the levels of [Y] expression.
CG17	increases expression	[X] treatment increases the levels of [Y] expression.
CG18	increases expression	[X] upregulates [Y] protein.
CG2	decreases activity	[X] results in decreased activity of [Y] protein.
CG21	increases phosphorylation	[X] results in increased phosphorylation of [Y] protein.
CG4	increases activity	[X] results in increased activity of [Y] protein.
CG6	decreases expression	[X] treatment decreases the levels of [Y] expression.
CG9	affects binding	[X] binds to [Y] protein.
CP1	decreases	[X] analog results in decreased phenotypes such as [Y].
CP2	increases	[X] induces phenotypes such as [Y].
CP3	affects	[X] affects phenotypes such as [Y].
GD1	marker/mechanism	Gene [X] is associated with diseases such as [Y].
GP1	association	Gene [X] is associated with pathways such as [Y].
UMLS		
UR116	clinically associated with	[X] is clinically associated with [Y].
UR124	may treat	The most widely used drug for preventing [X] is [Y].
UR148	due to	[X] induces [Y].
UR173	causative agent of	[X] is caused by [Y].
UR180	is finding of disease	[Y] is finding of disease [X].
UR196	has contraindicated class	[X] and [Y] has a drug-drug interaction.
UR211	biological process involves gene product	[X] involves [Y].
UR214	cause of	[Y] causes [X].
UR221	gene mapped to disease	[X] has a genetic association with [Y].
UR254	may be finding of disease	[X] has symptoms such as [Y].
UR256	may be molecular abnormality of disease	[Y] has a genetic association with [X].
UR44	may be prevented by	[X] treats [Y].
UR45	may be treated by	[X] treats [Y].
UR48	physiologic effect of	[X] results in [Y].
UR49	mechanism of action of	[X] has a mechanism of action of [Y].
UR50	therapeutic class of	[X] is a therapeutic class of [Y].
UR588	process involves gene	[X] involves [Y] process.
UR625	disease has associated gene	[X] has a genetic association with [Y].
UR97	contraindicated with disease	[X] has contraindicated drugs such as [Y].
Wikidata		
P2175	medical condition treated	[X] has effects on diseases such as [Y].
P2176	drug used for treatment	The standard treatment for patients with [X] is a drug such as [Y].
P2293	genetic association	Gene [X] has a genetic association with diseases such as [Y].
P4044	therapeutic area	[X] cures diseases such as [Y].
P780	symptoms	[X] has symptoms such as [Y].

Table 6: Manual prompts used in our experiments. Each prompt is created by domain experts.

Relation ID	Relation Name	IE	BioBERT		Bio-LM	
			Manual	Opti.	Manual	Opti.
CTD						
CD1	therapeutic	**14.29/22.33**	3.28/10.16	6.45/16.15	7.20/15.45	7.79/15.51
CD2	marker/mechanism	3.87/6.41	3.28/6.19	**9.81/23.60**	4.64/9.06	6.56/13.44
CG1	decreases expression	0.15/0.15	0.00/0.00	0.32/0.81	1.16/4.94	**4.59/7.99**
CG18	increases expression	**6.70/19.86**	0.00/0.71	0.00/0.19	0.94/8.58	1.46/7.52
CG2	decreases activity	**8.58/19.49**	0.00/0.00	4.81/8.29	0.67/2.90	4.08/13.41
CG21	increases phosphorylation	5.04/20.54	0.00/2.71	**14.73/18.14**	0.00/13.95	0.00/15.19
CG4	increases activity	7.83/**20.83**	0.00/0.00	0.83/2.69	0.00/0.89	**8.42**/20.59
CG9	affects binding	**10.88/23.13**	0.00/0.00	0.23/0.27	0.91/7.94	1.50/5.85
CP1	decreases	0.00/0.00	0.00/0.00	**6.37/19.33**	0.00/1.27	2.35/12.11
CP2	increases	0.00/0.00	0.00/0.14	**10.66/18.38**	0.00/0.81	0.00/0.14
CP3	affects	0.00/0.00	0.00/0.00	0.00/**1.51**	0.00/0.22	0.00/0.22
GD1	marker/mechanism	**4.17**/8.01	0.33/**11.64**	0.00/1.67	2.20/8.34	1.43/7.36
GP1	association	1.59/2.04	0.00/17.14	**16.69/31.67**	4.43/21.11	4.00/18.55
UMLS						
UR116	clinically associated with	6.35/**19.28**	0.84/4.07	**7.90**/18.08	2.64/11.02	6.13/14.87
UR124	may treat	**20.52/42.24**	1.03/4.31	1.76/4.10	10.69/25.35	8.79/22.76
UR173	causative agent of	0.31/0.31	1.41/4.84	9.81/**30.62**	4.53/15.63	**9.84**/27.78
UR180	is finding of disease	0.00/0.23	0.00/0.00	8.57/**29.72**	0.00/0.00	**9.63**/15.30
UR211	biological process involves gene product	0.00/0.00	0.49/1.85	**12.35**/24.08	0.00/0.25	9.30/**31.86**
UR214	cause of	1.74/2.44	0.00/1.92	3.83/7.80	1.05/7.32	**3.94/11.88**
UR221	gene mapped to disease	0.00/0.00	0.00/0.00	0.00/0.00	0.00/1.66	**14.44/30.27**
UR254	may be finding of disease	0.00/0.00	10.94/24.26	17.50/37.10	**27.71/38.41**	**27.71/38.41**
UR256	may be molecular abnormality of disease	0.00/0.33	0.00/0.00	0.00/0.00	0.00/0.00	**10.85/19.02**
UR44	may be prevented by	6.89/13.43	1.77/5.65	2.12/7.28	1.24/7.95	**8.83/20.71**
UR45	may be treated by	**17.10/26.22**	1.76/5.80	0.70/4.85	1.76/13.26	8.73/20.39
UR48	physiologic effect of	0.00/0.00	0.00/0.00	**3.06/7.47**	0.00/0.00	1.12/6.03
UR49	mechanism of action of	0.00/0.00	0.00/0.00	0.13/1.14	0.00/0.00	**1.17/3.64**
UR50	therapeutic class of	0.00/0.00	0.12/2.05	**7.17**/14.14	3.50/10.25	6.73/**21.98**
UR588	process involves gene	0.00/0.00	0.13/1.93	**4.66**/22.47	0.00/1.93	2.60/**29.73**
UR625	disease has associated gene	**3.56**/7.34	0.00/4.40	1.72/3.61	1.89/**9.02**	2.26/8.39
Wikidata						
P2175	medical condition treated	2.45/7.34	0.64/5.92	3.19/11.04	9.40/21.11	**9.47/24.94**
P2176	drug used for treatment	**22.82/47.24**	1.07/4.10	0.78/9.20	22.46/39.75	20.14/39.57
P2293	genetic association	**9.07/16.61**	0.00/7.77	1.04/4.38	2.24/11.43	2.90/9.21
P4044	therapeutic area	0.26/0.79	4.74/9.21	4.21/8.53	**9.47/19.47**	7.53/18.58
P780	symptoms	0.55/5.80	11.88/29.01	11.82/31.38	**16.30/37.85**	12.98/33.43

Table 7: Performance on each relation. Acc@1 and Acc@5 are reported. Best performances are in boldface.

Relation ID	Subject	Top 5 Predictions
CTD - CD1 (7.79 / 15.51)	[X] prevents diseases such as [Y].	
	Nitric Oxide	**Hypertension**, Multiple Sclerosis, Cardiac, Pulmonary, Cardiovascular
	Triamterene	**Hypertension**, Epilepsy, Diabetes, Cardiac, Myocardial
	SH-6 compound	Epilepsy, Cancer, Liver, Malignant, Inf
	quizartinib	Cancer, Liver, Hypertension, Leukemia, Sarcoma
CTD - CD2 (6.56 / 13.44)	[X] exposure is associated with significant increases in diseases such as [Y].	
	Normetanephrine	**Hypertension**, Cancer, Asthma, Hepatitis, Diabetes
	Vitamin K 1	**Hypertension**, Hepatitis, Cancer, Diabetes, Anemia
	lomefloxacin	Hypertension, Hepatitis, Cancer, Asthma, Diabetes
	cefditoren	Hypertension, Hepatitis, Diabetes, Cancer, Asthma
UMLS - UR173 (9.84 / 27.78)	[X] is caused by [Y] .	
	Meningococcal rash	**Meningococcus**, Streptococcus, Meningococci, Streptococcus pyogenes, Bacteria
	Macular syphilide	**Bacteria**, Virus, T. pallidum, **Treponema pallidum**, Legionella
	Braxy	Bacteria, Virus, Bacterial, Agents, Toxin
	Blister with infection	Virus, Adenovirus, Viral, Rotavirus, Enterovirus
UMLS - UR211 (9.30 / 31.86)	[X] involves [Y] .	
	Protein Kinase C	**Signaling**, Signal, Signal Processing, Apoptosis, Transcription
	Guanylate Cyclase	**Signaling**, Transcription, **Cell Signaling**, Calcium Signaling, Signal Processing
	HLA Complex	Transcription, Immune, Immune Response, Signal Processing, Infection
	gephyrin	signaling, Channel Regulation, Receptor Signaling, Signal Processing, . . .
UMLS - UR221 (14.44 / 30.27)	[X] has a genetic association with [Y] .	
	DICER1 syndrome	**DICER1 gene**, DICER, DICER gene, DICER1, DIC gene
	Cervical Wilms Tumor	**WT1 gene**, WT1, RET gene, PTEN gene, RET
	Gangliosidosis GM1	GM1 gene, GM1, GM gene, gene, GGM1 gene
	BALT lymphoma	BCL2 gene, ALT gene, BALT gene, ALK gene, ALK
UMLS - UR256 (10.85 / 19.02)	[Y] has a genetic association with [X] .	
	carcinosarcoma of lung	**TP53 Gene Inactivation**, TP53 Inactivation, RAS, **TP53 gene mutation**, EGFR
	Liver carcinoma	**TP53 Gene Inactivation**, TP53 Inactivation, KRAS Inactivation, KIT, RET
	Classical Glioblastoma	TP53 Inactivation, TP53 Gene Inactivation, EGFR, RET, MYC Gene Amplification
	Intratubular Seminoma	TP53 Inactivation, ERG, TP53 Gene Inactivation, KIT, KIT Inactivation
Wikidata - P2175 (9.47 / 24.94)	[X] has effects on diseases such as [Y].	
	amoxapine	**depression**, obsessive compulsive disorder, schizoaffective disorder, anxiety, . . .
	sofosbuvir	**chronic hepatitis C**, HIV, AIDS, HCV, hepatitis C virus
	duvelisib	AIDS, HIV, cancer, breast cancer, chronic obstructive pulmonary disease
	arsenic trioxide	AIDS, diabetes, cancer, tuberculosis, chronic obstructive pulmonary disease
Wikidata - P780 (12.98 / 33.43)	[X] has symptoms such as [Y].	
	legionnaires' disease	**fever**, pneumonia, fever and cough, cough and fever, **cough**
	Bocavirus infection	**fever**, conjunctivitis, jaundice, **diarrhea**, pneumonia
	pulmonary tuberculosis	hemoptysis, haemoptysis, dyspnea, cough, chest pain
	parenchymatous neurosyphilis	headache, fever, headache and fever, fever and headache, meningitis

Table 8: Top 5 predictions of Bio-LM (w/ OptiPrompt) given each prompt and different subjects. For each relation, we also report its Acc@1/Acc@5. Correct predictions are in boldface.