Semi-Supervised Training for Biomedical Question Answering

Dimitra Panou[1], Martin Reczko[1,*]

[1] Institute for Fundamental Biomedical Science, Biomedical Sciences Research Center "Alexander Fleming", 34 Fleming Street, 16672 Vari, Greece

Abstract

The recently introduced semi-supervised method GANBERT for finetuning large language models [1] has been applied for document relevance prediction in biomedical question answering. The additional use of unlabeled texts during training enhances the robustness of the prediction and outperforms our previous transformer ELECTROLBERT [2]. The initial document selection phase used both for ELECTROLBERT and GANBERT has been improved using BM25 combined with RM3 query expansion with optimized parameters. Both systems were continuously improved during the BioASQ11 [3] competition and in the last batch, GANBERT ranked as the 3rd team for document prediction. The previous version of ELECTROLBERT took the 1st place for the "yes/no" type questions in this years SYNERGY [4] prediction.

Keywords

Biomedical Question Answering, Semi-supervised learning, BioASQ, GANBERT, large language models

1. Introduction

One major bottleneck in the development of robust question answering systems is the lack of large volumes of high quality question answer pairs provided by human experts. Though transfer learning by finetuning pretrained large language models (LLMs) alleviates this problem [5], the limited data jeopardizes finetuning through overfitting. A recently suggested remedy [1] transfers the successful paradigm of semi-supervised learning used in Generative Adversarial Networks (GAN) for image processing [6] to the finetuning of LLMs. GANBERT extends the fine-tuning of BERT with unlabeled data using GAN framework, where a *generator*(G) is trained to produce samples of the internal BERT representation resembling the distribution over the unlabeled data, and a *discriminator*(D) that is trained to distinguish samples of the generator from the real instances. By generating only the internal representation of text, GANBERT avoids the generation of "fake" text instances. It is an effective semi-supervised method that can improve the generalization capability. Using vast amounts of unlabeled texts during training, the scope of the language model can be expanded to facilitate the use of alternative formulations for the same semantic content. In the original GANBERT paper [1] tests were performed for news topic classification, question conceptual class prediction, sentiment analysis and text genre

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

*Corresponding author.

panou@fleming.gr (D. Panou); reczko@fleming.gr (M. Reczko)
0000-0002-0005-8718 (M. Reczko)

classification. Two GANBERT variants were later successfully used for predicting he check-worthiness of potential fake news in tweets [7]. In [8], the noise generation in GANBERT was optimized for the task of discriminating correct paraphrases of Spanish texts. In the following we describe optimized document selection and the application of GANBERT for document relevance prediction in biomedical question answering in the BioASQ11 competition [9]. We also provide details for the additional predictions with our ELECTROLBERT algorithm [2] in the same competition.

2. BM25 and RM3 hyperparameter optimization

To identify documents relevant for a question, we replace the TF/IDF method with the widely used BM25 [10]. BM25 has two parameters $k1$ and b. $k1$ is intuitively related to the rate of increase in a document's score from matching an additional occurrence of a term, where smaller $k1$ provides a faster increase. The parameter b controls the extent of document-length normalisation. The search is combined with RM3 [11], a classic pseudo-relevance feedback based query expansion model, to find related concepts. RM3 has three parameters, *terms* is the number of query expansion terms, *docs* is the number of top-ranked documents to obtain the expansion terms and *qw* defines the weight of the original query. The efficient Python implementation in the package Pyserini is used [12]. A gridsearch on these parameters to optimize the mean average precision (MAP) of the top 10 returned documents for the BioASQ11 training set provided the values that were used in all four batches of BioASQ11. A random search optimizing the average MAP of the top 10 returned documents for the 240 questions in the first three batches of BioASQ11 indicates potential improvements. The optimized parameters shown in table 1 clearly outperform the default settings.

3. Training, validation and test data

For finetuning GANBERT, all pairs of a question and its correct documents provided in the training set for BioASQ11 are used for the 'relevant' class. As introduced in the ELECTROLBERT training [2], the negative examples for the 'non-relevant' class are generated using a range of false positives from the initial document selection phase to better discriminate the relevant documents obtained. All questions of the relevance training set were processed with BM25 and RM3 using the settings marked with B3+4:EB0-4 in table 1 to select 1000 relevant documents for each question. The documents were ranked according to their score and all documents between rank 100 and 150 were used as negative examples, excluding potential positive examples in these ranks. The values of the start and end rank positions for the negative set were optimized by retraining and maximizing the mean average precision measured on all batches of BioASQ10. For the unlabeled set, all pairs of a question and its ideal answer and all related snippets from the BioASQ10 training set were used. As a validation set, the top 100 documents scored with BM25 and RM3 (settings again as in B3+4:EB0-4) of the 240 questions in the first three batches of BioASQ11 was used. A final independent test was made on the 90 questions of batch 4 of BioASQ11.

Table 1
BM25 & RM3 parameter optimization. $k1$ & b are parameters of BM25 and the variables *terms* (Expansion terms), *docs* (number of top-ranked documents) and qw (Original query weight) are parameters of RM3 model. *nasnok* specifies the number of questions (total 240) with at least one correct document. *ndocok* specifies the number of correctly identified documents (max. 647). In the column "used for", Bx denotes the BioASQ11 test batch x, and EBy denotes the system ELECTOLBERTy.

$k1$	b	*terms*	*docs*	qw	MAP_{r123}	*nansok*	*ndocok*	used for
1.2	0.75	10	10	0.5	0.2734	128	192	defaults
0.4	0.3	10	10	0.5	0.2625	**138**	201	B1:EB2+3
1.1	0.0	10	10	0.5	0.2752	125	195	B1:EB0+1
0.4	0.3	17	14	0.6	0.2906	134	198	B2:EB0,2+3
0.30	0.31	16	16	0.8	0.2932	135	202	
0.40	0.31	20	16	0.7	0.2936	**138**	205	B3+4:EB0-4
0.40	0.31	20	16	0.9	0.2940	129	199	
0.30	0.31	20	16	0.8	0.2952	134	200	
0.45	0.36	20	21	0.8	0.2980	134	203	
0.40	0.31	20	16	0.8	0.2981	134	202	
0.60	0.37	17	16	0.8	0.2983	130	200	
0.50	0.33	20	25	0.7	0.2987	135	206	
0.45	0.37	15	22	0.7	0.2992	137	205	
0.45	0.31	17	20	0.7	0.2993	135	201	
0.40	0.38	15	24	0.8	0.2999	131	200	
0.40	0.30	18	21	0.7	0.3000	135	202	
0.55	0.34	19	23	0.7	0.3002	**138**	206	
0.60	0.34	14	18	0.8	0.3003	131	202	
0.35	0.34	18	25	0.7	0.3004	**138**	209	
0.35	0.37	17	26	0.7	**0.3011**	**138**	**210**	

4. GANBERT finetuning and hyperparameter optimization

The adaptation of the GANBERT architecture introduced in [1] for document relevance classification is shown in figure 1. Using the labeled and unlabeled data described in the previous section for finetuning and employing the large pretrained BERT model provided with the GANBERT implementation in the path for the real data (provided by the authors of GANBERT at https://github.com/crux82/ganbert), all relevant hyperparameters for GANBERT are optimized by multiple finetunings while monitoring the performance on the first three batches of BioASQ11 as shown in table 2. All GANBERT models perform substantially better when compared to the standard BERT model and the performance of GANBERT is quite stable for the different hyperparameter settings, also for variations as suggested in [8] in the noise generation part.

5. Results

In table 3 the performances of our document relevance submissions for the BioASQ11 competition are listed. All submissions marked with 'base model' use the ELECTROLBERT model of

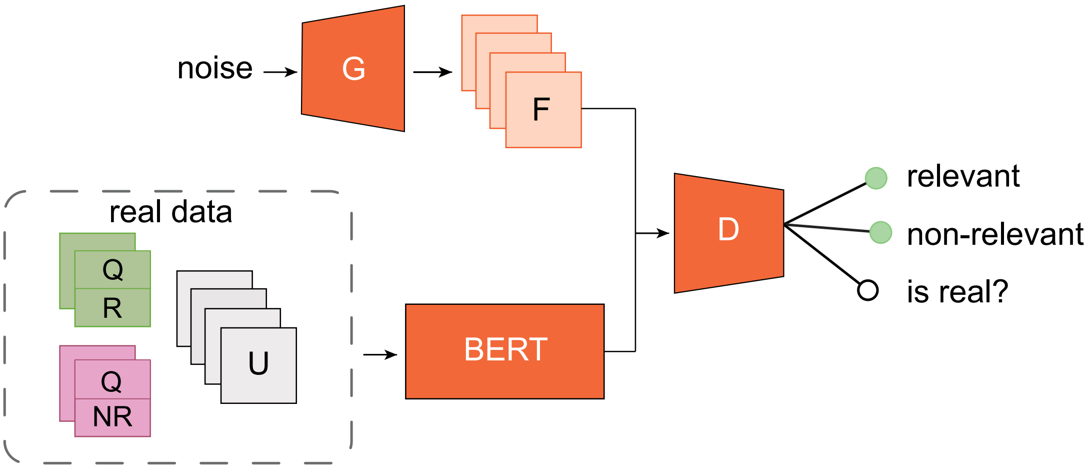

Figure 1: The GANBERT architecture for question answering: The *generator* G generates a set of fake representations F given a random distribution. These and the unlabeled U and labeled L vector representations computed by BERT are used as input for the *discriminator* D. The labeled examples are classified into documents relevant (R) and non-relevant (NR) for a question Q. The real data should be discriminated from the fake representations via the ’is real?’ output.

Table 2

Hyperparameter optimization for question answering GANBERT models using the MAP averaged over the first three batches of BioASQ11 (MAP_{b123}). The final test uses the MAP of BioASQ11 batch4 (MAP_{b4}). Unless specified, the sequence length for prediction is $SLEN_{predict} = 175$. All models are finetuned for 32000 steps. LR denotes the learning rate, $SLEN$ the sequence length during finetuning and $LABEL\_MASK$ controls the ratio between the number of labeled and unlabeled examples. The bold model is GANBERT3, submitted as ELECTOLBERT-4 in batch4.

Model	LR	SLEN	LABEL_MASK	Unlabeled set	Noise generation	MAP_{b123}	MAP_{b4}
BERT	$2e-6$	200	-	-	-	0.3166	0.2231
GANBERT	$\mathbf{2e-6}$	**200**	**0.02**	**BioASQ10**	**uniform[0,1]**	**0.3468**	0.2300
” , $SLEN_{predict} = 200$	”	”	”	”	”	0.3453	0.2223
” , $SLEN_{predict} = 150$	”	”	”	”	”	0.3456	0.2262
”	”	175	”	”	”	0.3459	0.2166
”	”	225	”	”	”	0.3363	0.2152
”	”	200	0.01	”	”	0.3395	**0.2301**
”	”	”	0.05	”	”	0.3380	0.2181
”	$1e-6$	”	0.02	”	”	0.3266	0.2066
”	$5e-6$	”	”	”	”	0.3334	0.2200
”	$2e-6$	”	”	BioASQ10+BoolQ [13]	”	0.3315	0.2166
”	”	”	”	BioASQ10	uniform[-1,1]	0.3333	0.2214
”	”	”	”	”	normal[0,1]	0.3289	0.2220

batch 4 in the BioASQ10 competition described in [2]. The models marked with 'large model' use the large architecture in [2], where pretraining was continued for 30 million steps and finetuning was performed with the labeled part of the training set for GANBERT for 180000 steps. It can be observed that the sequence length for predictions converged to an optimal value of $SLEN_{predict} = 175$ during the competition. With the optimized first phase document selection, it also became evident that the transformers in the second phase focus on the final ranking of the results and the number of documents was gradually reduced from $ndocs = 11500$ to $ndocs = 10$. In batch 4, the names of the systems *ELECTROLBERT-[1,2,3,4]* are used for the different GANBERT submissions.

Table 3
BioASQ11 document relevance prediction performance measured as mean average precision (MAP). The column 'model details' specifies the type of the transformer architecture, the sequence length during prediction and the number of documents to be ranked. The model GANBERT4 was trained for twice the number of steps as GANBERT3.

batch	MAP	system	per team rank	model details
1	**0.4590**	bioinfo-0	1	
	0.3875	ELECTROLBERT-2,3	4	base model, $SLEN_{predict} = 200$, $ndocs = 11500$
	0.3732	ELECTROLBERT-0,1	4	base model, $SLEN_{predict} = 250$, $ndocs = 11500$
2	**0.3852**	bioinfo-4	1	
	0.3252	ELECTROLBERT-2	4	base model, $SLEN_{predict} = 175$, $ndocs = 6750$
	0.2942	ELECTROLBERT-0	4	base model, $SLEN_{predict} = 250$, $ndocs = 6750$
	0.2781	ELECTROLBERT-3	4	base model, $SLEN_{predict} = 275$, $ndocs = 6750$
	0.2513	ELECTROLBERT-1	4	Query expansion using Roccio's methond [14]
				base model, $SLEN_{predict} = 400$, $ndocs = 300$
3	**0.3185**	dmiip2	1	
	0.2502	ELECTROLBERT-0	4	base model, $SLEN_{predict} = 175$, $ndocs = 60$
	0.2336	ELECTROLBERT-4	4	base model, $SLEN_{predict} = 175$, $ndocs = 16$
	0.2326	ELECTROLBERT-2	4	large model, $SLEN_{predict} = 200$, $ndocs = 13$
	0.2296	ELECTROLBERT-1	4	base model, $SLEN_{predict} = 150$, $ndocs = 300$
	0.2261	ELECTROLBERT-3	4	large model, $SLEN_{predict} = 150$, $ndocs = 11$
4	**0.3224**	dmiip3	1	
	0.2279	*ELECTROLBERT-1*	3	GANBERT4, $SLEN_{predict} = 175$, $ndocs = 10$
	0.2271	*ELECTROLBERT-4*	3	GANBERT3, $SLEN_{predict} = 175$, $ndocs = 10$
	0.2242	*ELECTROLBERT-3*	3	GANBERT2, $SLEN_{predict} = 175$, $ndocs = 11$
	0.2147	*ELECTROLBERT-2*	3	GANBERT1, $SLEN_{predict} = 175$, $ndocs = 16$
	0.1849	ELECTROLBERT-0	3	base model, $SLEN_{predict} = 175$, $ndocs = 60$

6. Conclusion and Future Work

Our suggested GANBERT version for document relevance prediction has shown promising performance, defeating our previous algorithm ELECTROLBERT. As can be seen at the published BioASQ11 results, both algorithms perform better than some of the other systems that seem to employ ChatGPT [15]. One obvious extension would be the replacement of BERT in the path processing the real data with ELECTROLBERT. This would also lead to the use of a more appropriate scientific vocabulary, as the BERT model provided with the GANBERT implementation uses a general purpose vocabulary. It should also be noted that the size of the unlabeled data set in this study is relatively small due to generation of this using only text available with the BioASQ datasets and our limited computational resources. One way to increase this could be the use of random segments from Pubmed abstracts.

Acknowledgments

GPU computations were offered by HYPATIA, the Cloud infrastructure of the Greek ELIXIR node.

References

[1] D. Croce, G. Castellucci, R. Basili, GAN-BERT: Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 2114–2119. URL: https://aclanthology.org/2020.acl-main.191. doi:10.18653/v1/2020.acl-main.191.

[2] M. Reczko, ELECTROLBERT: Combining Replaced Token Detection and Sentence Order Prediction, in: Proc. of CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy, online http://ceur-ws.org/Vol-3180/paper-24.pdf, urn:nbn:de:0074-3180-7, 2022.

[3] A. Nentidis, G. Katsimpras, A. Krithara, S. Lima-López, E. Farré-Maduell, L. Gasco, M. Krallinger, G. Paliouras, Overview of BioASQ 2023: The eleventh BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), 2023.

[4] A. Nentidis, G. Katsimpras, A. Krithara, G. Paliouras, Overview of BioASQ Tasks 11b and Synergy11 in CLEF2023, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018. URL: https://arxiv.org/abs/1810.04805. doi:10.48550/ARXIV.1810.04805.

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Nets, in: Proceedings of the 27th International Confer-

ence on Neural Information Processing Systems - Volume 2, NIPS'14, MIT Press, Cambridge, MA, USA, 2014, p. 2672–2680.

[7] R. A. Frick, I. Vogel, I. N. Grieser, Fraunhofer SIT at CheckThat! 2022: Semi-Supervised Ensemble Classification for Detecting Check-Worthy Tweets, Working Notes of CLEF (2022).

[8] T. Ta, A. B. S. Rahman, L. Najjar, A. Gelbukh, GAN-BERT, an Adversarial Learning Architecture for Paraphrase Identification, in: Proc. of IberLEF 2022, September 2022, A Coruña, Spain, 2022.

[9] A. Krithara, A. Nentidis, K. Bougiatiotis, G. Paliouras, BioASQ-QA: A manually curated corpus for Biomedical Question Answering, Scientific Data 10 (2023) 170.

[10] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, Experimentation as a way of life: Okapi at TREC, Inf. Process. Manag. 36 (2000) 95–108.

[11] V. Lavrenko, W. B. Croft, Relevance Based Language Models, in: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '01, Association for Computing Machinery, New York, NY, USA, 2001, p. 120–127. URL: https://doi.org/10.1145/383952.383972. doi:10.1145/383952.383972.

[12] J. Lin, X. Ma, S. Lin, J. Yang, R. Pradeep, R. F. Nogueira, Pyserini: An Easy-to-Use Python Toolkit to Support Replicable IR Research with Sparse and Dense Representations, CoRR abs/2102.10073 (2021). URL: https://arxiv.org/abs/2102.10073. arXiv:2102.10073.

[13] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, K. Toutanova, BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions, in: NAACL, 2019.

[14] V. Lavrenko, W. B. Croft, Relevance-based language models, in: ACM SIGIR Forum, volume 51, ACM New York, NY, USA, 2017, pp. 260–267.

[15] OpenAI, ChatGPT [Large language model] https://chat.openai.com/chat, 2023.