

ELECTROLBERT: Combining Replaced Token Detection and Sentence Order Prediction

Martin Reczko¹

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

Abstract

ELECTROLBERT is a novel transformer algorithm that combines the replaced token prediction of the ELECTRA system [1] with the sentence order prediction used in the ALBERT system [2]. As reported, the default next sentence prediction component used in most BERT-based transformers has the drawback that a random next sentence can be easily predicted based on its different scope. The sentence order prediction facilitates the detection of semantic flow and is well suited for finetuning question answering systems, as the pairing of a question with text related to the correct answer resembles the correct order of two sentences in a scientific text. The implementation of ELECTROLBERT is based on the BioELECTRA [3] code. ELECTROLBERT is pretrained on the 2022 baseline set of all PubMed abstracts provided by the National Library of Medicine and two predictors are finetuned using pairs of relevant and non-relevant question-abstract pairs for document prediction and examples for the “yes/no” type questions, both generated using the BioASQ10 training dataset [4]. For each novel question in the document prediction task of BioASQ10, 6750 Pubmed abstracts are filtered for processing by ELECTROLBERT from all Pubmed abstracts using a combination of the GENSIM topic modelling system [5] and a simple infrequent word detection method. The system was continuously improved during the BioASQ10 competition and in the last batch, ELECTROLBERT ranked as the 3rd team for document prediction and 1st place and team for the “yes/no” type questions.

Keywords

Biomedical Question Answering, ELECTRA, ALBERT, BioASQ

1. Introduction

Autoencoders with attention mechanisms led to impressive improvements in many natural language processing tasks [6]. A computationally efficient approach is called ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) [1] which replaces the traditional reconstruction of masked tokens from their context as introduced in BERT [7] with a simpler identification of replaced tokens that are generated by an adaptive generator. Another successful algorithm is the ALBERT (A Lite BERT) transformer that incorporates several parameter reduction techniques and replaces the next sentence prediction (NSP) component of BERT with a sentence order prediction (SOP) component that improves the detection of inter-sentence coherence. In the transformer called ELECTROLBERT (ELECTRA-A-LBERT) introduced here,

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

✉ reczko@fleming.gr (M. Reczko)

ORCID [0000-0002-0005-8718](https://orcid.org/0000-0002-0005-8718) (M. Reczko)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

Table 1

Question–Answer pairs in real abstracts. Coherence detected by sentence order prediction (SOP) facilitates answer identification. The topic-shift detection obtained with next sentence prediction (NSP) is less specific.

PMID	Question Answer
34884907	Neurogenic Inflammation in the Context of Endometriosis-What Do We Know? Endometriosis (EM) is an estrogen-dependent disease characterized by the presence of epithelial, stromal, and smooth muscle cells outside the uterine cavity.
34894155	The question to ask is, is this prescribed load regimen congruent with Wolff’s law, and does it provide an adequate mechanical stimulus to maintain the functional health of periodontal complex? This question was answered by studying the effects of mice chewing on soft food (SF) and hard food (HF) while undergoing experimental tooth movement (ETM).
34893939	Will Artificial Intelligence (AI) re-humanize or de-humanize medicine? As AI becomes pervasive in clinical medicine, we argue that the ethical framework that sustains a responsible implementation of such technologies should be reconsidered.

the SOP component of ALBERT is combined with the ELECTRA approach that is missing any component to predict inter-sentence relations.

2. Motivation

NSP is defined as a prediction if two sentences are consecutive in the training texts. The negative case uses sentences from two different documents. During the design of ALBERT, it was shown that a BERT-style NSP essentially fails on the SOP task, while SOP as used in ALBERT also has reasonable performance on the NSP problem. The conclusion was that NSP solves the easier task of predicting the topic-shift between the two segments and fails to model the discourse-level coherence between consecutive sentences. As shown in Table 1, it is not uncommon that scientific texts contain consecutive segments with a question and an answer. An efficient SOP would directly identify these cases as correct answers and finetuning a question-document relevance prediction could effectively reuse the parts of the embedding that the SOP developed during training. The embedding obtained with NSP based on detecting topic-shift trends to accept all answers containing a topic similar to the question, but do not necessarily answer the question. An overview of the system architecture is shown in Figure 1.

3. Data generation and training

3.1. Initial document retrieval

As the complete corpus cannot be processed by the transformer for each question in reasonable time, a computationally efficient filter has to filter the corpus according to each question. To this end, the TF/IDF based free topic modelling system GENSIM [5] (with logarithmic term frequency

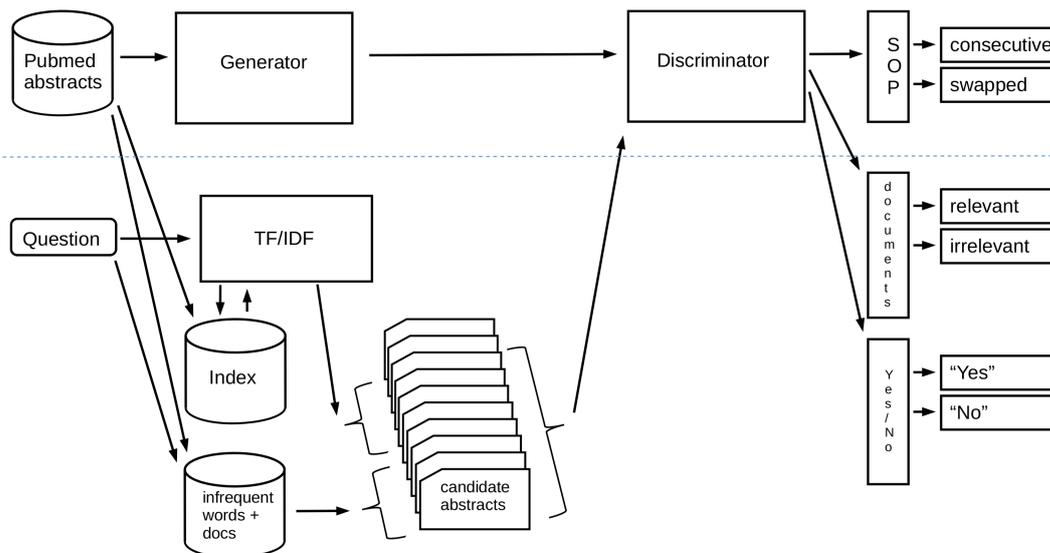


Figure 1: The ELECTROLBERT system. The components above the blue line are used during pretraining. During question answering, all components except the Generator are used.

weighting, IDF document frequency weighting and no document normalization) is used to return a list of 6750 documents relevant for each question to be processed by ELECTROLBERT. A question in the fifth batch was "What is Waylivra?" (id:626aea6ae764a5320400003d) with the answer in the document with PMID 3130103: "Volanesorsen (Waylivra®), an antisense oligonucleotide inhibitor of ...". As the special 'registered trademark' character "®" was included in the dictionary entry for this word, but not in the question, the document was not detected in my submission. A new dictionary and index was constructed omitting the special characters ®, © and ™.

3.2. Infrequent words treatment

Processing the corpus of 22542347 pubmed abstracts and titles, a dictionary with 5440825 words is obtained. To ensure the detection of infrequent words that would not be detected using the TF/IDF score, a list of documents is constructed for each word occurring less than 4000 times in the complete corpus. These words cover 99.55% of the dictionary. In case a question contains one of these infrequent words, the union of the lists of documents for each infrequent word occurring in the question is appended to the list of documents detected by TF/IDF for evaluation by ELECTROLBERT.

3.3. Pretraining data

Pretraining data was generated using a modified version of the ALBERT code. To emphasize sentence order prediction and subsequent question answering, all examples have two segments, where sentences longer than half of the maximal sequence length without separators are truncated to this length. In 50% of the examples, the sentence order is swapped. Instead of a

random 10% fraction of the examples having only a single segment, no single segment examples are used. The vocabulary available in BioELECTRA [3] containing 31620 entries¹ is used and all training texts are lowercase. The titles and abstracts are extracted from the pubmed baseline xml files with a custom R script based on the xml2 package.

3.4. Pretraining

For subsequent relevance prediction, a model with the 'base' configuration of BioELECTRA is pretrained due to computational limitations. This corresponds to an embedding size of 768, a generator hidden size of 256 and 12 hidden layers. The model was trained for 1.5 million steps using the ADAM algorithm ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-5}$, `learning_rate` = 10^{-4} , `weight_decay_rate` = 0.005, `max_seq_length` = 128, `train_batch_size` = 14). The final replaced token discriminator has an area under the the curve (AUC) of 0.92 ± 0.0087 . For finetuning the "Yes/No" task, a 'large' configuration is pretrained for 5 million steps with an embedding size of 1024, a generator hidden size of 256 and 24 hidden layers ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-5}$, `learning_rate` = 10^{-4} , `weight_decay_rate` = 0.0025, `max_seq_length` = 128, `train_batch_size` = 12). It should be noted that the training of the 'large' model did not yet converge after 5 million steps, with an discriminator AUC of 0.85 ± 0.001 . GPU memory restrictions required the smaller `train_batch_size` for the 'large' configuration. As known, smaller `batch_sizes` lead to less accurate gradient calculations. This, in combination with the larger number of parameters to optimize for the 'large' configuration ('large': 335 million vs 'base': 110 million) could explain the slow convergence.

3.5. Document relevance prediction finetuning

As initial question answering is focused on the "yes/no" type questions, 1148 questions of this type are extracted from the BioASQ10 training set (873 "yes", 275 "no"). All questions with shared answer documents are grouped into equivalence sets. Initially, all documents relevant for a questions are used as positive relevance examples and 2% of all documents not contained in the equivalence set for a question are used as negative relevance examples, leading to an approximate negative to positive example ratio of 20 to 1. As negative examples are generated from completely unrelated questions, the class boundaries are distant. To better discriminate the relevant documents obtained with GENSIM topic modelling, construction of the negative example was modified after the third batch of BioASQ10. All questions of the relevance training set were processed with GENSIM to produce 1000 documents for each question. The documents were ranked according to their TF/IDF score and all documents between rank 750 and 875 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 was optimized by retraining and maximizing the mean average precision measured on the second batch.

The model was trained for 18768 steps using the ADAM algorithm ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-5}$, `learning_rate` = $2 \cdot 10^{-6}$, `weight_decay_rate` = 0.003, `max_seq_length`

¹https://github.com/SciCrunch/bio_electra/blob/master/electra/data/pmc_2017_abstracts_wp_vocab_sorted.txt

Table 2

BioASQ10 prediction performance: Document relevance and “yes/no” task. For documents, the mean average precision (MAP) is used for evaluation.

batch	documents				“yes/no” type questions	
	BioASQ submission MAP	per team rank	final system MAP	per team rank	accuracy	per team rank
1	0.1121	7	0.3649	5	-	-
2	0.1632	9	0.3090	5	-	-
3	0.3209	8	0.3666	6	0.76	10
4	0.3101	6	0.3140	6	0.75	11
5	0.3242	4	0.3242	4	0.6429	10
6 ²	0.0977	3	0.0977	3	1.0	1

= 200, train_batch_size = 16).

3.6. “Yes/No” prediction finetuning

A training set of 70% of the “Yes/No” type questions was extracted from the BioASQ10 training set. To prevent trivial tests, the training set, the 10% validation set and the 20% test set all contain questions that are not in the equivalence set of any question in all other sets. The model was trained for 88260 steps using the ADAM algorithm ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-5}$, learning_rate = $4 \cdot 10^{-5}$, weight_decay_rate = 0.008, max_seq_length = 256, train_batch_size = 4). To classify a question, all documents relevant for the question are scored by ELECTROLBERT. If the average score over all documents exceeds a fixed threshold, the answer to the question is “Yes”. In the opposite case, an additional requirement to assign “No” to a question is a certain degree of unanimity in the scores of all relevant documents. Only if the variance of all scores is below a fixed threshold, the final answer is negative.

3.7. Fast abstract retrieval

A custom R script accessing chunks of 100000 abstracts based on their PMIDs can retrieve and store > 3000 abstracts with titles per minute using 15 CPUs. This step is one of the major bottlenecks to guarantee the processing of 100 questions within 24 hours without the use of large compute clusters.

4. BioASQ10 results

The performance of ELECTROLBERT for each batch of BioASQ10 is listed in Table 2. For the document retrieval task, the submissions that reflect the different development stages are listed as well as a retrospective analysis for batches 1 to 4 that is obtained with the final system used in batch 5 and 6. The improvements during the development of the system are obvious.

²Batch 6 consisted of questions posed by new biomedical experts interested in material and answers that can be automatically provided by state-of-the-art IR and QA systems. It is not part of the official evaluation.

5. Conclusions and Next Steps

All document retrieval results of ELECTROLBERT have been obtained with the 'base' sized architecture and are thus promising, fueling the expectation of competitive performance once training of the 'large' architecture has converged. ELECTROLBERT will be finetuned for the snippet, factoid and list tasks. Transfer learning from SOP to relevance prediction and from relevance prediction to the "Yes/No" task will be evaluated.

Acknowledgments

GPU computations were offered by HYPATIA³, the Cloud infrastructure that supports the computational needs of the Greek ELIXIR community.

References

- [1] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, 2020. URL: <https://arxiv.org/abs/2003.10555>. doi:10.48550/ARXIV.2003.10555.
- [2] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, 2019. URL: <https://arxiv.org/abs/1909.11942>. doi:10.48550/ARXIV.1909.11942.
- [3] K. r. Kanakarajan, B. Kundumani, M. Sankarasubbu, BioELECTRA:pretrained biomedical text encoder using discriminators, in: Proceedings of the 20th Workshop on Biomedical Language Processing, Association for Computational Linguistics, Online, 2021, pp. 143–154. URL: <https://aclanthology.org/2021.bionlp-1.16>. doi:10.18653/v1/2021.bionlp-1.16.
- [4] A. Nentidis, G. Katsimpras, E. Vandorou, A. Krithara, L. Gasco, M. Krallinger, G. Paliouras, Overview of bioasq 2021: The ninth bioasq challenge on large-scale biomedical semantic indexing and question answering, Experimental IR Meets Multilinguality, Multimodality, and Interaction (2021) 239–263. URL: http://dx.doi.org/10.1007/978-3-030-85251-1_18. doi:10.1007/978-3-030-85251-1_18.
- [5] R. Rehurek, P. Sojka, Gensim–python framework for vector space modelling, NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic 3 (2011).
- [6] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, 2020. URL: <https://arxiv.org/abs/2006.03654>. doi:10.48550/ARXIV.2006.03654.
- [7] 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.

³<https://hypatia.athenarc.gr>