ANGEL: Enterprise Search System for the Non-Profit Industry

Saiful Haq*, **Ashutosh Sharma***, **Pushpak Bhattacharyya**
Department of Computer Science and Engineering, IIT Bombay, India
saifulhaq@cse.iitb.ac.in, sharma96@illinois.edu, pb@cse.iitb.ac.in

Abstract

Non-profit industry need a system for accurately matching fund-seekers (*e.g.* AMERICAN NATIONAL RED CROSS) with fund-givers (*e.g.*, BILL AND MELINDA GATES FOUNDATION) aligned in cause (*e.g.*, cancer) and target beneficiary group (*e.g.*, children). In this paper, we create an enterprise search system "ANGEL" for the non-profit industry that takes a fund-giver's mission description as input and returns a ranked list of fund-seekers as output, and vice-versa. ANGEL employs ColBERT, a neural information retrieval model, which we enhance by exploiting the two techniques of (a) Syntax-aware local attention (SLA) to combine syntactic information in the mission description with multi-head self-attention and (b) Dense Pseudo Relevance Feedback (DPRF) for augmentation of short mission descriptions. We create a mapping dictionary to curate a "non-profit-search database" containing information on **594K** fund-givers and **194K** fund-seekers from IRS-990 filings for the non-profit industry search engines . We also curate a "non-profit-evaluation" dataset containing scored matching between 463 fund-givers and 100 fund-seekers. The research is in collaboration with a philanthropic startup that identifies itself as an "AI matching platform, fundraising assistant, and philanthropy search base." Domain experts at the philanthropic startup annotate the non-profit evaluation dataset and continuously evaluate the performance of ANGEL. ANGEL achieves an improvement of **0.14** MAP@10 and **0.16** MRR@10 over the state-of-the-art baseline on the non-profit evaluation dataset. To the best of our knowledge, ours is the first effort at building an enterprise search engine based on neural information retrieval for the non-profit industry.

1 Introduction

Non-profit industry consists of non-profit foundations (NPFs), non-profit service providers (NPSE), and independent donor individuals. NPFs and independent donor individuals, who function as fund-givers (*e.g.*, BILL AND MELINDA GATES FOUNDATION), supply funds to NPSEs, who function as fund-seekers (*e.g* AMERICAN NATIONAL RED CROSS). A significant number of financial transactions occur between fund-givers and fund-seekers. In 2020, Americans gave $ 471.44 billion to charity. Among all charitable giving, independent donor individuals contributed 69%, NPFs contributed 19%, and corporations contributed the remaining 4% (Hadero, 2021).

Fund-seekers employ "donor research analysts" that raise funds from fund-givers that are aligned in the philanthropic cause (*e.g.*, cancer) and target beneficiary group (*e.g.*, children). The median salary of a "donor research analyst" in the United States is $ 50K. Small-scale fund-seekers cannot hire "donor research analysts" due to the lack of budget, and the absence of an exhaustive search of aligned fund-givers makes small-scale fund-seekers repeatedly seek funds from known and prominent fund-givers. Similarly, due to lack of time, fund-givers often find fund-seekers using personal connections and donate without understanding the impact they can create by exhaustively searching and donating to relevant fund-seekers.

We create an enterprise search system "ANGEL", in collaboration with a philanthropic startup, that can accurately match fund-givers with fund-seekers using publicly available data on non-profit organizations in IRS-990 (Internal revenue service) filings [*]. ANGEL reduces the overheads related to "donor research analysts" compensation for fund-seeking organizations and search time for individual donors.

The IRS-990 filing of a non-profit organization (fund-giver or fund-seeker) consists of multiple forms, each containing multiple fields. Some fields

[*]Equal Contribution

[*]https://www.irs.gov/charities-non-profits/form-990-series-downloads

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 828–835
December 6-10, 2023

provide information on the organization's philanthropic cause, which is necessary for matching fund-givers with fund-seekers, and others provide information on the organization's finances used for filtering it based on spending. We create a mapping dictionary "non-profit-dict" for IRS-990 data to accurately map similar fields to a common field and drop irrelevant fields to match fund-givers with fund-receivers. In this way, we curate a non-profit-search database for the non-profit industry search engines containing information on **594K** fund-givers and **194K** fund-seekers.

Enterprise search engines use Information Retrieval (IR) models that take a query as input and return a ranked list of relevant documents from the search database. Textual IR has witnessed the use of language models to obtain contextual vector representations of queries and documents for vector-based matching instead of keyword-based matching. Neural information retrieval models have shown considerable improvement in accuracy compared to keyword-based information retrieval models like BM25 (Robertson et al., 2009). To our knowledge, existing enterprise search engines (e.g. Propublica non-profit explorer) for the non-profit industry do not use neural information retrieval models.

Enterprise search engines based on neural information retrieval (IR) employ models such as ColBERT (Khattab and Zaharia, 2020) (Santhanam et al., 2021), that uses BERT encoder (Devlin et al., 2018) to obtain contextualized token vectors representing mission descriptions of fund-seekers and matches them with the contextualized token vectors of grant descriptions of fund-givers. We augment the capability of ColBERT for the task of enterprise search in non-profit industry by exploiting the two techniques of (a) Syntax-aware Local Attention (SLA) and (b) Dense Pseudo Relevance Feedback (DPRF). We train our models on MSMARCO passage ranking dataset (Bajaj et al., 2016). To evaluate the search quality of ANGEL, we curate a non-profit-evaluation dataset from the non-profit-search database. The dataset is a 463 by 100 matrix, where each row represents a fund-giver and each column represents a fund-seeker. Each index in the 2-dimensional matrix is a matching score ranging from 0 to 9. Domain experts from the philanthropic startup have annotated this dataset to evaluate ANGEL. We compare the performance of ANGEL with ColBERTv2 (Santhanam et al., 2021) on the non-profit-evaluation dataset. Through ablation study, we observe that ANGEL based on ColBERT performs better than ColBERT-v2, a more potent IR model, on the non-profit-evaluation dataset.

Our contributions are:

- Non-profit-dict, a mapping dictionary to map 400 variables parsed from IRS-990 Filings using IRSx[*] python package to 17 relevant variables with the objective of matching fund-givers with fund-seekers. To the best of our knowledge, this is the first time such a dictionary has been created to curate data from IRS-990 Filings.

- Syntax-aware Local Attention (SLA) using dependency parsing for improving retrieval accuracy of IR systems. To the best of our knowledge, this is the first work utilizing SLA-augmented contextual token vectors for information retrieval. SLA-ColBERT achieves an improvement of **0.02** MAP@10 and **0.03** MRR@10 over ColBERTv2 on the non-profit evaluation dataset.

- ANGEL, an enterprise search system based on ColBERT that uses Syntax-aware Local Attention (SLA) and Dense Pseudo Relevance Feedback (DPRF). ANGEL achieves an improvement of **0.14** MAP@10 and **0.16** MRR@10 over the state-of-the-art model ColBERT-v2 on the non-profit-evaluation dataset. To the best of our knowledge, this is the first neural IR system for the non-profit industry, thereby providing a solid baseline for the task of enterprise search in the non-profit industry.

2 Related work

2.1 Enterprise search system for Non-Profit industry

Many keyword-based enterprise search engines (e.g., Propublica non-profit explorer) on the IRS 990 database have emerged after the release of the IRS 990 Electronic filing data in 2016. Such keyword-based search engines do not consider the context in which the words are used in the input query and documents while performing retrieval and re-ranking. ANGEL uses contextualized token vectors obtained using the BERT Encoder. These vectors also consider the context in which words are used in the input query and documents present

[*] https://github.com/jsfenfen/990-xml-reader

in the corpus while performing retrieval and re-ranking.

2.2 Neural information retrieval

IR has witnessed the use of large language models to obtain contextual vector representations of query and document tokens. These vectors are further used for retrieval and ranking. For an input query, a candidate list of documents is first retrieved using approximate K nearest neighbor search on embedding indexes (Johnson et al., 2019). The retrieved documents are re-ranked using neural re-rankers (Nogueira and Cho, 2019). We can classify neural IR models into interaction-focused and representation-focused models based on the type of neural re-rankers (Khattab and Zaharia, 2020). In interaction-focused models, a query document pair is given to a cross-encoder as an input. In representation-focused models, query and document are given separately to a dual encoder. The output of the encoder in a neural IR model is passed through a score aggregation function to produce a relevance score. The relevance score quantifies the relevance of a document for a given input query and is used to rank the retrieved documents. Representation-focused models are faster than interaction-focused systems as they precompute the document vectors offline. ColBERT and ColBERT-v2 (Santhanam et al., 2021) are BERT-based representation-focused IR systems that use a low-cost max-sim operator based on cosine similarity to calculate a document's relevance score for a given input query. This score is used for re-ranking. COIL (Gao et al., 2021) integrates lexical matching with contextualized vector-based scoring. The contextualized vectors are precomputed for all documents and stored in an inverted-index format. At inference time, cosine-based scoring is done only for lexical matched tokens in the query and the document. This reduces computation cost attributed to nearest neighbor search while keeping the benefits of contextualized vector-based scoring. Polyencoder (Humeau et al., 2019) integrates cross-attention-based scoring with representation-focused models. Polyencoder generates and stores sentence vectors of documents offline. At inference time, Polyencoder computes attention between query tokens and trained codecs to generate vectors which are further passed through an attention layer with document vector to generate the final document vector. The final document and sentence vector are used to score the query-document pair. ANGEL uses ColBERT with two techniques discussed in section 4.3 and 4.2.

2.3 Syntax guided self-attention

In recent years, dependency grammar has been incorporated in transformers' architecture to improve results on downstream tasks like named entity recognition and machine translation. (Zhang et al., 2020) introduced a self-attention layer after the transformer encoder, in which the tokens are allowed to attend to their ancestors in the dependency parse tree. (Strubell et al., 2018) limited the self-attention of one attention head in the transformer using a dependency parse tree. Within this attention head, each token can only attend to its syntactic parents. (Li et al., 2020) used a dependency parse tree to generate a masking matrix at each layer of the transformer encoder. This matrix is used to prevent distant tokens from attending in self-attention.

We propose to use Syntax-aware Local Attention using dependency parsing (Li et al., 2020) to improve the encoder performance in ColBERT. The details for the technique is discussed in section 4.3. Compared to previous works, this is the first time that dependency parsing based information is utilized for the task of information retrieval.

2.4 Dense Pseudo Relevance Feedback

Pseudo-relevance feedback (Abdul-Jaleel et al., 2004) (Amati, 2003) uses statistical information like the term frequency of the retrieved document tokens to augment the input query with relevant document tokens. Dense pseudo-relevance feedback uses vector operations (e.g., clustering and cosine similarity) to augment the input query with relevant document tokens. (Diaz et al., 2016) (Kuzi et al., 2016) used token vectors closest to input query token vectors in the static word vector space for query expansion. (Zheng et al., 2020) used contextualized vectors to select document chunks closest to the documents in the pseudo-relevance feedback set and re-rank the documents using the chunk vectors. In (Wang et al., 2021), a query is given as an input to the ColBERT model, and top N documents are retrieved based on their relevance scores. K-means clustering is performed on token vectors present in the retrieved documents to find K cluster centroids that can represent the retrieved document vector space. The token vectors closest to the cluster centroids and the Inverse Document

Frequency (IDF) values of the corresponding tokens in the document corpus is identified. The cluster centroid vectors weighted by the IDF scores corresponding to the closest token vectors are appended to the input query vectors, and the new augmented query is then used to retrieve the final set of ranked documents.

3 Dataset

We curate a "non-profit-search" database and a human-annotated "non-profit-evaluation" dataset in collaboration with a philanthropic startup to test the system's capability of matching fund-givers with fund receivers. In this section, we first give details about the IRS-990 filings used to populate our dataset. After that, we discuss the variable mapping created to curate our non-profit-search database efficiently. Once the non-profit-search database is populated, a subset of it is given to the philanthropic startup for annotation that results in the non-profit-evaluation dataset. At the end of this section, we discuss the strategy for annotation.

3.1 IRS 990 Dataset

United states Internal Revenue Service (IRS) mandates non-profit organizations to file a tax return called IRS-Form-990 every year. IRS-Form-990 provides the public with information about a Non-Profit. This information includes the non profit's operating location, finances, mission statement, activities, executive names, executive salaries etc. There are other variations of IRS-Form-990: IRS-Form-990-PF, IRS-Form-990-EZ and IRS-Form-990-N. Fund-givers file IRS-Form-PF irrespective of their financial status. Fund-receivers file IRS-Form-990, IRS-Form-990-EZ and IRS-Form-990-N. Fund-recievers with gross receipts more than $200,000, or total assets more than $500,000 file Form 990. Fund receivers with gross receipts less than $200,000, and total assets less than $500,000 file IRS-Form-990-EZ. IRS-Form-990-EZ is an abbreviated four-page version of IRS-Form-990. Fund-receivers with annual gross receipts less than $50,000 do not have to file the complete IRS-Form-990 (although they can opt to do so). Instead, they may file the IRS-Form-990-N, also called the "e-Postcard".

In 2016, IRS released the electronic version of Form 990 and its variations. This reduced the time-consuming and costly process of converting paper records to digital records via manual data entry or Optical Character Recognition (OCR). The electronic version is the main source of data for our non-profit-search database and non-profit-evaluation dataset. It contains data of fund-seekers and fund-givers in extensible markup language (XML) files.

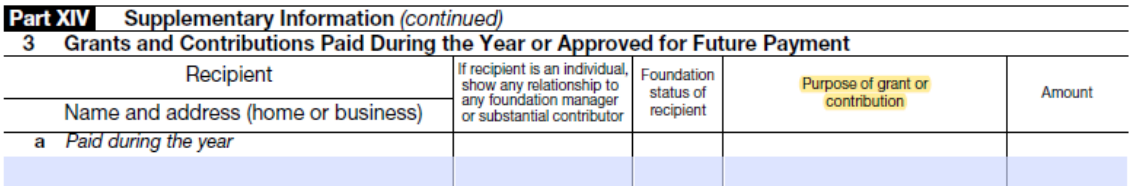
3.2 Mapping dictionary

Despite the fact that the IRS has made the data available, it is inaccessible due to its complex extensible Markup Language (XML) structures. We have used IRSx[*], a python package to collect and parse IRS-990 filings. The parsed data is stored in more than 100 variables that represent fields like organization name, city, state, name of grantee organization, purpose of grants, grant size, asset size, mission statements, number of employees, employee name, employee salary, employee designation, etc.

An IRS filing consists of Form 990 or one of its variants followed by schedules. These schedules are represented by letters A, B, C, D, E, F, G, H, I, J, K, L, O and R. Form 990 and its variant may contain similar fields that are named differently. For the purpose of matching fund-giver with a fund-receiver we need to find the vector representation of the interest area of both kinds of organizations.

As shown in Figure 1, the field "Briefly describe the organizations mission or most significant activities" in Part 1 "Summary" of "Form 990" describes the interest area of the fund-receiver. There is another field "Briefly describe the organization's mission" in Part 3 "Statement of Program Service Accomplishments" of "Form 990" that represents the interest area of a fund-receiver. Similarly, there are fields in each of the schedules that are indirectly related to the interest area of the organization. There is a need to merge text data from similar variables to a common variable using a mapping dictionary. The IRS corpus contains data on organizations that have filed Form 990 or its variants with different schedules. There is a need for a pipeline that can parse relevant data irrespective of the type of filing and stores it to a local storage. We have analysed the official IRS documentation and developed a variable mapping dictionary "non-profit-dict" that can map similar variables to a common variable irrespective of the type of filing and document. This dictionary can map 400 variables present in Form 990, its variants and various schedules to 17 relevant variables. After parsing, we obtain a single

[*]https://github.com/jsfenfen/990-xml-reader

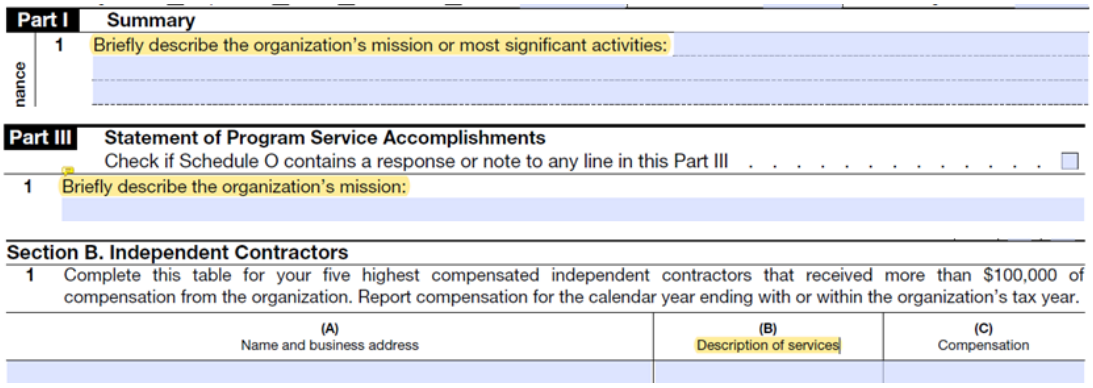


Part XIV Supplementary Information *(continued)*

3 Grants and Contributions Paid During the Year or Approved for Future Payment

Recipient Name and address (home or business)	If recipient is an individual, show any relationship to any foundation manager or substantial contributor	Foundation status of recipient	Purpose of grant or contribution	Amount
a *Paid during the year*				

(a)



Part I Summary

1 Briefly describe the organization's mission or most significant activities:

nance

Part III Statement of Program Service Accomplishments

Check if Schedule O contains a response or note to any line in this Part III . . . . . ☐

1 Briefly describe the organization's mission:

Section B. Independent Contractors

1 Complete this table for your five highest compensated independent contractors that received more than $100,000 of compensation from the organization. Report compensation for the calendar year ending with or within the organization's tax year.

(A) Name and business address	(B) Description of services	(C) Compensation

(b)

Figure 1: Sample of Form 990-PF (a) and Form 990 (b). The fields highlighted in yellow represent interest area of a Non-Profit organization and are mapped to common variable using the mapping dictionary

table having **788K rows** and **17 columns**. Each row in the table represents a non-profit organization and each column represents a feature (e.g asset size, interest area, expenses, revenue, website address, phone number, etc.) of the non-profit organization. This table represents the non-profit-search database. The data summary is discussed in Table 2 .

3.3 Annotation process

We have identified a subset of organizations located in "New York" and working in "education" domain from the "non-profit-search" database. The document corpus, which contains 100 documents, is made using the column "interest area" in the curated "non-profit-search" database with a filter on "fund-receivers". The set of queries, which contains 463 queries, is made using the column "interest area" in the curated "non-profit-search" database with a filter on "fund-givers". For every query, all the documents are given relevancy labels based on mission statements on the scale of 0 to 9. Since each document is labelled for each query, the annotation is complete as opposed to MSMARCO dataset, which contains sparse judgements. The dataset has been annotated by two domain experts at the non-profit philanthropic startup and is continuously being evaluated and expanded. The inter-annotator agreement (Cohen-Kappa) between them is found to be 0.473.

4 Methodology

4.1 ColBERT

We use ColBERT (Khattab and Zaharia, 2020) as the base architecture for ANGEL. The model architecture as shown in figure 2, comprises of (a) a query encoder, (b) a document encoder, and (c) a late interaction mechanism. Given a query with q tokens and a document with d tokens, the Query encoder obtains q fix sized token embeddings, and the document encoder obtains d fix sized token embeddings. The maximum input sequence length for the query, q_{max}, and, for the document, d_{max}, is set before giving them to the respective encoders. If q is less than q_{max}, we append $q_{max} - q$ tokens to the input query, and if q is greater than q_{max}, q is truncated to q_{max}. If d is less than d_{max}, it is kept as it is with no padding. If d is greater than d_{max}, d is truncated to d_{max}.

4.2 Dense pseudo-relevance feedback

We propose to use a modified version of (Wang et al., 2021) to improve the encoder performance in ColBERT. Instead of selecting K cluster centroid vectors as feedback token vectors and appending them to the input query vector, K tokens closest to

K cluster centroids in the euclidean space are selected as feedback tokens. We append the feedback tokens to the query and generate the query vector using the expanded query. To determine how well the feedback tokens discriminate the document collection, top M tokens are selected based on their Inverse-document-frequency scores. These M tokens are added back to the input query and final set of retrieved documents are obtained by performing retrieval with the expanded query.

4.3 Syntax-aware Local Attention for encoding of documents

We use Syntax-aware Local Attention (SLA) (Li et al., 2020) to enhance token level embedding for information retrieval task. We obtain dependency mask for Query and documents with one or more sentences. The dependency graph in the latter case will be a collection of isolated graphs, where each graph represents a sentence. We obtain dependency graph of a query or document and treat it like an undirected graph. In the graph, each token x_i is mapped to a tree node v_i, and the path length between node v_i and v_j in the graph is denoted by $dis\left(v_i, v_j\right)$. The distance $D(i,j)$ between token x_i and x_j present in the same sentence is given as:

$$D(i,j) = \min dis\left(v_k, v_j\right), \quad k \in [i-1, i+1]$$

The distance between token x_i and x_j present in different sentences is given as:

$$D(i,j) = \infty$$

Then, in order to determine whether token x_j can attend to token x_i, a threshold m is applied to restrict the distance $D(i,j)$. The mask matrix $\mathbf{M}^{loc}$ is formulated as:

$$\mathbf{M}_{ij}^{loc} = \begin{cases} 0, & D(i,j) \leq m \\ -\infty, & otherwise \end{cases}$$

Given the query $\mathbf{Q}$ and key $\mathbf{K}$ projected from the hidden vectors $\mathbf{H}$, the syntax-aware local attention scores $\mathbf{S}^{loc}$ are formally defined as:

$$\mathbf{S}^{loc} = softmax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}} + \mathbf{M}^{loc}\right)$$

where d is the hidden dimension of query and key matrices. In this local attention, two tokens can attend to each other only if they are close enough in the dependency tree.

Model	MAP@10	MRR@10
ColBERTv2(baseline)	0.33	0.34
SLA-ColBERT	0.35	0.37
DPRF-ColBERT	0.33	0.35
ANGEL	**0.47**	**0.50**

Table 1: Results on the non-profit-evaluation Dataset. The models in bold are improvements done to the baseline model as part of this work. ANGEL (SLA-DPRF-ColBERT) performs better than the baseline ColBERTv2 model.

5 Experiments

We compare the performance of ANGEL (SLA-DPRF-ColBERT) model with the official "ColBERTv2 checkpoint"[*], which has been trained for retrieval for a significantly higher number of iterations (approximately 200k). We discuss the model training configurations in detail in B. For evaluating the performance of our models, we use the non-profit-evaluation dataset. The dataset contains 463 queries and 100 documents. Each document has a relevance score on a scale of 0 to 9. The threshold to classify a document as relevant is decided empirically. The average number of relevant documents per query for a threshold of 2 is 47.44. The average number of relevant documents per query for a threshold of 3 is 16.72. The average number of relevant documents per query for a threshold of 4 is 4.76. We selected a threshold of 3 as it neither gives too many relevant documents, nor too less relevant documents per query.

6 Results

Our results show 44.5% gain in MAP@10 and 49.7% gain in MRR@10 score over the baseline on the non-profit-evaluation dataset from using Dense-psuedo-relevance-feedback and syntax-aware-self-attention in conjunction. The results for these experiments are presented in Table 1. To better understand the importance of each of the techniques for the accuracy gain, we compare them in detail in A.

7 Summary, conclusion, and future work

In this paper, we show the motivation behind creating an enterprise search engine "ANGEL" for the non-profit domain. Training ANGEL on the multilingual version of the MSMARCO dataset with the IR objective, to support non-English queries and documents remains as future work.

[*] https://github.com/stanford-futuredata/ColBERT

References

Nasreen Abdul-Jaleel, James Allan, W Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark D Smucker, and Courtney Wade. 2004. Umass at trec 2004: Novelty and hard. *Computer Science Department Faculty Publication Series*, page 189.

Giambattista Amati. 2003. *Probability models for information retrieval based on divergence from randomness Ph. D*. Ph.D. thesis, thesis. University of Glasgow.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fernando Diaz, Bhaskar Mitra, and Nick Craswell. 2016. Query expansion with locally-trained word embeddings. *arXiv preprint arXiv:1605.07891*.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. Coil: Revisit exact lexical match in information retrieval with contextualized inverted list. *arXiv preprint arXiv:2104.07186*.

Haleluya Hadero. 2021. Americans gave a record $471 billion to charity in 2020$.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Saar Kuzi, Anna Shtok, and Oren Kurland. 2016. Query expansion using word embeddings. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 1929–1932.

Zhongli Li, Qingyu Zhou, Chao Li, Ke Xu, and Yunbo Cao. 2020. Improving bert with syntax-aware local attention. *arXiv preprint arXiv:2012.15150*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2021. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *arXiv preprint arXiv:2112.01488*.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium. Association for Computational Linguistics.

Xiao Wang, Craig Macdonald, Nicola Tonellotto, and Iadh Ounis. 2021. Pseudo-relevance feedback for multiple representation dense retrieval. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 297–306.

Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. 2020. Sg-net: Syntax-guided machine reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9636–9643.

Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, and Andrew Yates. 2020. Bert-qe: contextualized query expansion for document re-ranking. *arXiv preprint arXiv:2009.07258*.

A Enterprise search results

In this section, we analyze search results of enterprise search or information retrieval models for input query "girl education". For analysis, we look at Top-2 search results.

A.1 Propublica Non-profit explorer

Propublica non-profit explorer is an enterprise search engine over IRS-990 electronic filings and scanned PDFs of raw filings. We use the advanced full-text search option where it returns documents with any mention of query terms in the document's body.

The search engine operates in two modes : Normal and Boolean. In normal mode, the search engine returns "10" organizations for the input query "girl education". Top-2 search results are two separate IRS-990 filings of the non-profit BRIGHT PINK NFP in 2012 and 2013. Top-2 output is not relevant as mission description of "BRIGHT PINK NFP" is related to "breast and ovarian cancer in women" as mentioned in Part 1 of IRS-990 filing

made in 2022. In Boolean mode, the search engine returns 1,939,007 results for the input query "girl education". The Top-2 results are "AARON AND CATIE ENRICO FAMILY FOUNDATION" with IRS-990-PF filing and "AFCEA EDUCATIONAL FOUNDATION" with IRS-990 filing. Both of these non-profit organizations contain "0" mentions of "girl" or related terms in their IRS-990 Filings.

On manually going through the search database over which Propublica non-profit explorer works, we found a relevant non-profit organization EDUCATE GIRLS with the following text "To promote and support girl education by facilitating community involvement and responsibility for local school reform" in Part 1 of its IRS-990 filing made in 2023.

Propublica non-profit explorer fails to perform accurate full-text search over its search database as the top-2 search results are not relevant to the input query "girl education".

A.2 ColBERTv2

In this section, ColBERTv2 is used as non-profit search engine. First, ColBERTv2 store's the non-profit evaluation dataset in FAISS embedding index. After the storage is complete, ColBERTv2 perform's search using the input query "girl education". The non-profit-evaluation dataset is a test subset of the non-profit search database and it only contains selected few filings from the IRS-990 electronic filings. It can be used for quick evaluation of the search quality.

The Top-2 results are non-profits "ST. LUKE's SCHOOLS" with mission description "a coeducational episcopal day school, preschool through eighth grade, for students of all faiths" in Part 1 of its IRS-990 filing made in 2022 and "NEW YORK WOMENS FOUNDATION" with mission description "create an equitable and just future for women and families across new york city" in Part 1 of its IRS-990 filing made in 2022 .

ColBERTv2 fails to perform an accurate full-text search as "NEW YORK WOMENS FOUNDATION" focuses majorly on women, not girls.

A.3 ANGEL

In this section, ANGEL is used as the non-profit search engine. Like the ColBERTv2-based search engine described in A.2, ANGEL first stores the non-profit evaluation dataset in the FAISS embedding index and then performs a search using the input query "girl education".

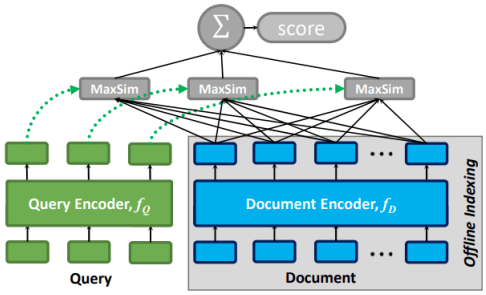


Figure 2: ColBERT architecture for encoding of queries and documents.

Description	Value
No. of non-profits	788k
No. of fund-givers	594k
No. of fund-seekers	194k
Feature per non-profit	17

Table 2: Non-profit-search database summary.

The Top-2 results are non-profits "ST. LUKE's SCHOOLS" with mission description "a coeducational episcopal day school, preschool through eighth grade, for students of all faiths" in Part 1 of its IRS-990 filing made in 2022 and " RUDOLF STEINER SCHOOL INC" with mission description "rudolf steiner school embraces waldorf education, a pedagogy derived from the insights of rudolf steiner" in Part 1 of its IRS-990 filing made in 2023.

ANGEL performs accurate full-text search over its search database as the top-2 search results are relevant to the input query "girl education".

B Experiments

We train SLA-ColBERT on the MSMARCO passage ranking dataset to reduce the triplet loss objective. The dataset contains 8.8M documents, 532k query-relevant document pairs, and 39 million triplets. We train the model for 20k iterations with a batch size of 128 on the first 2.56 million training triplets, each triplet $< q, d_+, d_- >$ containing a query q, a positive passage d_+ and a negative passage d_-. We also employ in-batch negatives per GPU, where a cross-entropy loss is applied to the positive score of each query against all passages corresponding to other queries in the same batch. The BERT encoder is finetuned from the official "bert-base-uncased checkpoint" and the remaining parameters are trained from scratch.