M3: A Multi-Task Mixed-Objective Learning Framework for Open-Domain Multi-Hop Dense Sentence Retrieval

Yang Bai, Anthony Colas, Christan Grant, Daisy Zhe Wang
The University of Florida
Gainesville, Florida, USA
{baiyang94, acolas1, christan, daisyw}@ufl.edu

Abstract

In recent research, contrastive learning has proven to be a highly effective method for representation learning and is widely used for dense retrieval. However, we identify that relying solely on contrastive learning can lead to suboptimal retrieval performance. On the other hand, despite many retrieval datasets supporting various learning objectives beyond contrastive learning, combining them efficiently in multi-task learning scenarios can be challenging. In this paper, we introduce **M3**, an advanced recursive **M**ulti-hop dense sentence retrieval system built upon a novel **M**ulti-task **M**ixed-objective approach for dense text representation learning, addressing the aforementioned challenges. Our approach yields state-of-the-art performance on a large-scale open-domain fact verification benchmark dataset, FEVER. Code and data are available at: https://github.com/TonyBY/M3

Keywords: Information Retrieval, Fact Verification, Contrastive Learning, Multi-task Learning, Multi-hop Retrieval

1. Introduction

Open-domain fact verification (Thorne et al., 2018; Jiang et al., 2020; Bai et al., 2023) is a challenging task where single-hop or multi-hop sentence-level evidence for a given claim needs to be extracted from a large pool of documents to verify human-generated claims (See Figure 1 for an example from the FEVER dataset). A three-stage approach is commonly used to solve the problem (see Figure 2) (Thorne et al., 2018; Yoneda et al., 2018; Hanselowski et al., 2018; Nie et al., 2019a; Zhong et al., 2019; Zhou et al., 2019; Soleimani et al., 2020; Subramanian and Lee, 2020; Jiang et al., 2021; Krishna et al., 2022; Fajcik et al., 2023; DeHaven and Scott, 2023). In the first step, the retriever produces a list of n candidate documents given a claim. From the top-n documents, a sentence reranker selects the top-k sentences. Lastly, a claim classifier predicts the claim verdict based on the top-n sentences.

Claim: **Sheryl Lee** has yet to appear in a ***film*** as of ***2016***.
Evidence Documents
Doc1: Sheryl Lee *In 2016*, she appeared in **Café Society**, and also completed the Showtime revival of Twin Peaks (2017), reprising her role of Laura Palmer. **Doc2: Café Society** *Café Society is a* 2016 American romantic comedy-drama *film* written and directed by Woody Allen .
Verdict: Refuted

Figure 1: A FEVER example where multi-hop sentence-level evidence from multiple Wikipedia documents is required for verification.

Retrieval models have traditionally relied on term-based information retrieval (IR) methods (Thorne et al., 2018; Yoneda et al., 2018; Nie et al., 2019a; Jiang et al., 2021; DeHaven and Scott, 2023), which do not capture the semantics of a claim beyond lexical matching and remain a key bottleneck. In contrast, recent works train neural network-based encoders to obtain dense representations of queries and documents in vector spaces and then use maximum inner-product search (MIPS) to complete the retrieval (Karpukhin et al., 2020; Khattab and Zaharia, 2020; Qu et al., 2020; Ren et al., 2021; Xiong et al., 2021a,b; Wu et al., 2021; Zhang et al., 2022). When compared to traditional IR approaches, these dense retrievers have demonstrated significant improvements.

There are, however, some issues with current dense information retrieval models. First of all, they are trained on datasets at the document/passage level. This can increase the possibility of learning suboptimal representations due to internal representation conflicts (Wu et al., 2021). In particular, a passage can be organized by multiple semantically different sentences. It is not optimal to model a passage like this as a unified dense vector. Moreover, each document/passage consists of multiple sentences, from which multiple semantically distant queries (questions, claims, etc.) can be derived. In contrastive learning frameworks, that can benefit from large batch sizes when doing in-batch negative sampling, such a one-to-many problem can lead to severe conflicts when two conflicting

LREC-COLING 2024, pages 10846–10857
20-25 May, 2024.

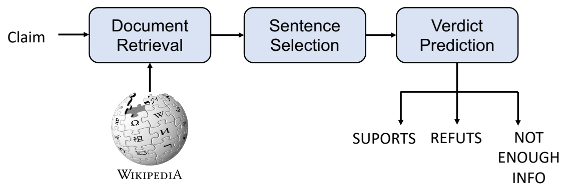

Figure 2: Canonical thee-stage fact verification framework.

claims with the same context document/passage are sampled within the same batch (Wu et al., 2021). Hence, to avoid the above-mentioned conflicts, we propose to use dense sentence-level retrieval as the first-level retriever to replace the traditional document retrieval modules in the canonical open-domain fact verification pipeline.

Furthermore, we observe that current dense information retrieval models rely solely on contrastive objectives, which could prevent the models from learning better representations and subsequently result in suboptimal recall. On the other hand, despite many retrieval datasets supporting various learning objectives beyond contrastive learning, combining them efficiently in multi-task learning scenarios can be challenging. In this paper, we introduce **M3**, an advanced recursive **M**ulti-hop dense sentence retrieval system built upon a novel **M**ulti-task **M**ixed-objective approach for dense text representation learning, addressing the aforementioned challenges. Our approach yields state-of-the-art performance on a large-scale open-domain fact verification benchmark dataset, FEVER.

Contributions of this paper include:

- We present an advanced recursive multi-hop dense sentence retrieval system (M3) based on a novel dense sentence representation learning method, which achieves state-of-the-art multi-hop retrieval performance on the FEVER dataset.
- We propose a novel dense sentence representation learning method (M3-DSR) based on multi-task learning and mixed-objective learning frameworks that significantly outperforms strong baselines such as BM25 (Yang et al., 2017) and DPR (Karpukhin et al., 2020) on sentence-level retrieval.
- We introduce an efficient heuristic hybrid ranking algorithm for combining retrieved single-hop and multi-hop sentence evidence, which shows substantial improvements over previous methods.
- We developed an end-to-end multi-hop fact verification system based on M3 that achieves state-of-the-art performance on the FEVER dataset.

2. Background and Related Works

2.1. Dense Text Retrieval

Dense Text Retrieval (DTR) (Karpukhin et al., 2020; Khattab and Zaharia, 2020; Xiong et al., 2021a; Wu et al., 2021; Zhang et al., 2022) has gained significant attention in recent years due to its potential to revolutionize sparse retrieval methods such as TF-IDF (Ramos, 2003) and BM25 (Robertson et al., 1994; Yang et al., 2017) on the document retrieval task. Contrastive learning, at its core (Hadsell et al., 2006; Reimers and Gurevych, 2019; Karpukhin et al., 2020), aims to learn effective representations by contrasting similar and dissimilar pairs of data. In the context of DTR, this translates to training models to distinguish between relevant and non-relevant document-query pairs.

In recent studies, Akkalyoncu Yilmaz et al. (2019) and Wu et al. (2021) propose to improve dense passage retrieval based on sentence-level evidence. In particular, Wu et al. (2021) investigated contrastive conflicts in the contrastive learning framework when performing document/passage-level representation learning as discussed in Section 1.

In contrast, we propose a simple approach that bypasses such conflicts by performing dense sentence-level retrieval in combination with multi-task, mixed-objective learning that shows stronger empirical performance.

2.2. Multi-hop Text Retrieval

The multi-hop text retrieval method is crucial to complex question-answering (Nie et al., 2019b; Xiong et al., 2021b; Li et al., 2023) and complex fact-verification (Thorne et al., 2018; Yoneda et al., 2018; Hanselowski et al., 2018; Nie et al., 2019a; Jiang et al., 2021; DeHaven and Scott, 2023) tasks where evidence is aggregated from multiple documents before logical reasoning or multi-hop reasoning is applied to infer the answer or a verdict. In prevailing approaches (Nie et al., 2019a; Asai et al., 2019), a document graph is constructed based on entity linking or hyperlinks found in the underlying Wikipedia corpus. These methods, however, might not be generalizable to new domains, where entity linking might perform poorly, or hyperlinks might be sparse(Xiong et al., 2021b).

With a recursive framework, MDR (Xiong et al., 2021b) applies dense retrieval to the multi-hop setting. Utilizing efficient MIPS methods, it iteratively encodes the question and previously retrieved documents as a query vector and retrieves the next relevant documents. Aly and Vlachos (2022) propose a retrieve-and-rerank method, AdMIRaL, consisting of a retriever that jointly scores documents in the knowledge source and sentences from previously retrieved documents and achieves the state-

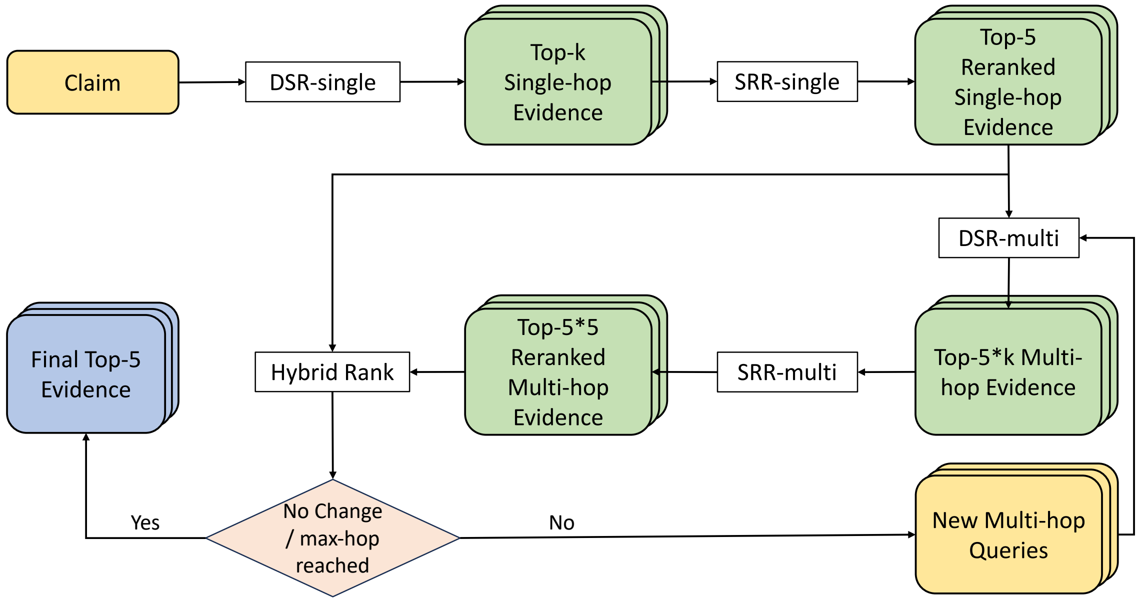

Figure 3: M3 iterative dense sentence retrieval pipeline. DSR refers to the dense sentence retrieval model; SRR refers to the sentence reranking model; *-single and *-multi indicate whether the model is trained on single-hop or multi-hop examples. When no specific number of hops is given, the multi-hop retrieval process continues until the top-5 hybrid-ranked sentences stop changing.

of-the-art multi-hop document retrieval recall on the FEVER dataset. However, to ensure efficiency, AdMIRaL's first-stage retriever uses sparse retrieval (i.e., BM25), which may sacrifice retrieval recall.

In our work, we combine the advantages of MDR and AdMIRaL to develop M3 for dense multi-hop search using a recursive retrieve-and-rerank framework. We also propose a hybrid ranking algorithm to jointly rank the single-hop and multi-hop retrieval results and achieve better overall retrieval recall.

Moreover, unlike MDR and AdMIRaL, which focus only on the retrieval of multi-hop document-level evidence at the first stage, M3 is capable of achieving state-of-the-art retrieval performance for both sentence-level and document-level evidence, which is more challenging and offers more fine-grained evidence that is crucial for downstream multi-hop inference. Our fact-verification system based on M3 achieves the highest claim classification accuracy on the blind FEVER testing dataset.

3. Method

3.1. Overview

Our work focuses on improving the retrieval component of open-domain fact verification. With a claim c in natural language and a collection of M text documents, the retrieval module needs to retrieve a ranked list of sentence-level evidence $S : \{s_1, s_2, ..., s_k\}(k << M)$ that provides sufficient information for downstream logical inference components to determine whether the claim c is supported, refuted, or unverifiable by the facts in the corpus. It is important to note that M can be very large (for example, in our setting, there are over 5 million documents with over 25 million sentences), and k should be small ($k = 5$ in the FEVER setting).

Our multi-hop dense sentence retriever M3 uses an iterative sentence-level retrieve-and-rerank scheme to recursively retrieve evidence (see Figure 3). The sentence retrieval probability at each step depends on the previous retrievals, i.e., $P(s_t|c, s_1, ..., s_{t-1})$. In practice, this probability is calculated as $P(s_t|q_{t-1})$ where $q_{t-1} = c \oplus s_1 \oplus ... \oplus s_{t-1}$, and $\oplus$ refers to concatenation operator. When $t = 1$, the retrieval probability is only conditioned on the original claim.

M3 differs from the existing multi-hop dense document retrieval method(Xiong et al., 2021b) in four ways: 1) finer retrieval granularity: document-level -> sentence-level, 2) we add reranking after each step of retrieval, and 3) when combining single-hop and multi-hop retrievals, a novel hybrid ranking algorithm is used, 4) we train a novel dual-encoder model using multi-task and mixed-objective learning to learn better dense text representations that yield higher retrieval recalls.

3.2. Dense Sentence Retrieval

Dense sentence retrieval aims to learn low-dimensional and continuous representation for the queries and sentences in the corpus in order to efficiently retrieve the top-k sentence-level evidence through an approximate nearest neighbor (ANN)

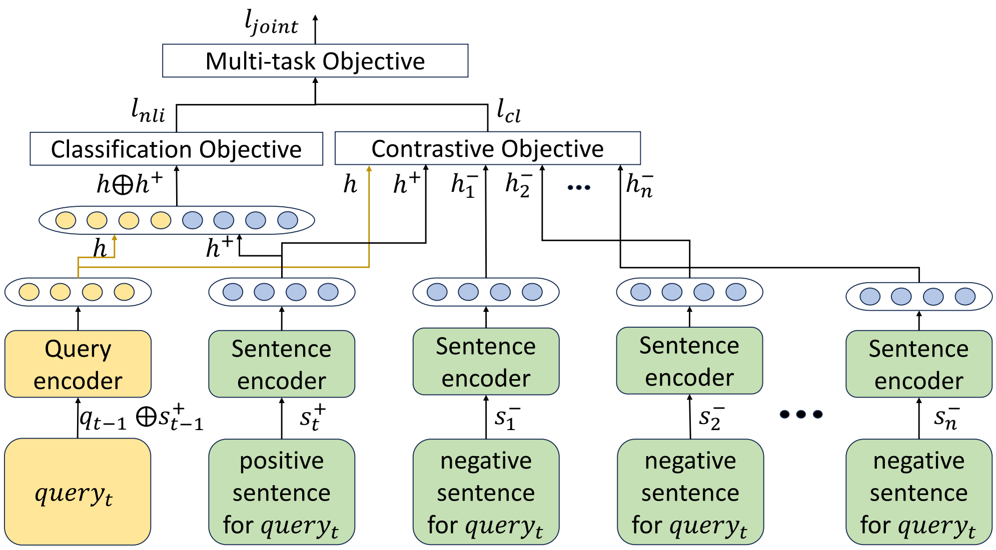

Figure 4: M3-DSR multi-task learning framework. When $t = 1$ (i.e., first-hop), the input of the query encoder is the original claim c.

(Johnson et al., 2017) method.

We train our dense sentence retrievers (M3-DSR) with a novel multi-task mixed-objective learning method. Using this method, we are able to learn better sentence representations that yield better retrieval recall (see more ablation studies Section 6).

3.2.1. Multi-task Learning

In the FEVER dataset, evidence and verdict annotations are given for each claim. Naturally, we explore training better text encoders with the FEVER dataset through multi-task learning with two objectives, contrastive and (claim) classification. Figure 4 shows an overview of our multi-task learning framework.

Contrastive Learning Objective The contrastive objective is implemented as in (Karpukhin et al., 2020; Gao et al., 2021), where each input query x_i is paired with a positive example x_i^+ and n negative examples $\{x_{i,1}^-, x_{i,2}^-, ..., x_{i,n}^-\}$. We also use the in-batch negative sampling (Karpukhin et al., 2020) taking other examples in the same batch as "negatives", and the model predicts the positive one among negatives to approximate the softmax over all examples. Let $\mathbf{h}_i$ and $\mathbf{h}_i^+$ denote the representations of x_i and x_i^+, the training objective ℓ_i is then defined as:

$$\ell_{cl\_i} = -\log \frac{e^{\text{sim}(\mathbf{h}_i,\mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N}\left(e^{\text{sim}(\mathbf{h}_i,\mathbf{h}_j^+)/\tau} + e^{\text{sim}(\mathbf{h}_i,\mathbf{h}_j^-)/\tau}\right)}. \quad (1)$$

where τ is a temperature hyperparameter, N is batch size and $\text{sim}(\mathbf{h}_1, \mathbf{h}_2)$ is the inner product $\mathbf{h}_1^\top \mathbf{h}_2$. The input sentences are encoded by a transformer language model: $\mathbf{h} = f_\theta(x)$. Specifically, we obtain $\mathbf{h}$ for a special token $[CLS]$, which represents the whole input sequence.

When sampling negative examples, we follow (Karpukhin et al., 2020) using BM25 (Lin et al., 2021) to retrieve top sentences from the whole corpus that are not included in the evidence annotation set. In addition, we filter these top negative examples with a more complex attention-based model to eliminate those that are too close to the claims to avoid including too many false negatives. Specifically, we use an off-the-shelf pre-trained sentence ranker that scores sentences based on their semantic similarity to the query. An empirical threshold is set based on extensive observation. We show that reducing the false negative examples in the training data is crucial to contrastive learning (see Section 6 for more details).

Classification Objective The encoded query $\mathbf{h_i}$ and positive example $\mathbf{h_i^+}$ is used to calculate the claim label probability $P(y|(\mathbf{h_i}, \mathbf{h_i^+}))$:

$$P(y|(\mathbf{h_i}, \mathbf{h_i^+})) = softmax_y(Linear(\mathbf{h_i} \oplus \mathbf{h_i^+})). \quad (2)$$

where $\oplus$ refers to concatenation operator.
The claim classification (NLI) loss is then defined as:

$$\ell_{nli\_i} = CrossEntropy(y^*, P(y|(\mathbf{h_i}, \mathbf{h_i^+}))). \quad (3)$$

Multi-Task Objective Our multi-task objective is a linear combination of contrastive loss and classification loss:

$$\ell_{joint\_i} = \alpha * \ell_{cl\_i} + \beta * \ell_{nli\_i}. \quad (4)$$

where hyperparameters: α and β are $\in [0, 1]$

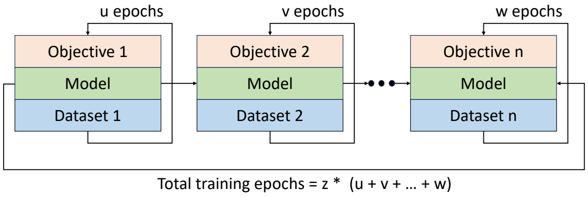

Figure 5: M3-DSR mixed-objective learning framework. The same model is trained with different dataset-objective combinations sequentially.

3.2.2. Mixed-Objective Learning

Many datasets are proposed for training encoders for open-domain dense text retrieval (Kwiatkowski et al., 2019; Joshi et al., 2017; Berant et al., 2013; Baudis and Sedivý, 2015; Rajpurkar et al., 2016; Thorne et al., 2018), however, they may support different objectives (e.g., contrastive, question answering, natural language inference, or multi-task). To make maximum use of these valuable datasets, we developed a framework that allows dense encoders to be trained over these datasets with different user-defined objectives and intervals/epochs (see Figure 5 for more details). The framework is not trivial as it allows us to optimize the model across multiple datasets that support different objectives more flexibly and conveniently with access to hyperparameters such as the objective and appearance frequency of each dataset during training. In extensive experiments, this framework demonstrated its effectiveness in improving retrieval recall.

3.3. Sentence Reranking

In sentence reranking, the top retrieved sentences from the previous step are ranked again using a more sophisticated method. Following (DeHaven and Scott, 2023), we reformulate this task as a sentence pair classification task: given a query-sentence pair (q, s), predict labels from {SUPPORTS, REFUTES, NOT ENOUGH INFO (NEI)}. The relevancy score for sentence ranking is calculated as:

$$Score(s) = 1 - softmax_{NEI}(h_{q \oplus s}) \tag{5}$$

where $h_{q \oplus s}$ is the embedding of a concatenated query-sentence pair encoded by a language model. $sofmax_{NEI}$ calculates the normalized confidence score over the NEI (irrelevant) class.

From the top retrievals in the last step, we sample the negative (NEI) sentences to create the training data (see Section 4 for more details).

3.4. Hybrid Ranking

In FEVER, not all claims have multi-hop evidence. To maximize the overall retrieval recall, we propose a dynamic hybrid ranking algorithm to jointly rank the single-hop and multi-hop retrievals. Inspired by (Ma et al., 2021) who explored and demonstrated effective methods of combining retrieval results from dense and sparse retrievers through a simple normalization and linear combination, we demonstrate this idea also works when combining single-hop and multi-hop retrievals. In addition, we scale the retrieval score for each step of multi-hop retrieval through production. This step is important because it ensures that each episode of evidence is proportional to the other. A detailed implementation is presented in Algorithm 1.

Algorithm 1 Hybrid Ranking Algorithm.

Require: (1) $ScoreMap_{single}$, a dictionary that saves the top single-hop retrievals in key-value pairs, e.g., $\{\cdots, se_i : sc_i, \cdots\}$, where key (se_i) is a sentence id, and value (sc_i) is the corresponding score acquired by Equation 5;
(2) $SequenceList$, a list of top multi-hop retrieval paths that consist of t id-score pairs, each pair representing one step of iterative retrieval results of t-hops, e.g., $[\cdots, \underbrace{((se_i, sc_i), (se_j, sc_j), \cdots, (se_k, sc_k))}_{t}, \cdots]$.
(3) mth and $\gamma \in (0, 1]$, hyperparameters that need to be tuned.

function HYBRID_RANK($ScoreMap_{single}$, $SequenceList$, mth, γ)
  $ScoreMap_{multi} = \{\}$
  for seq in $SequenceList$ **do**
    seq_score = *Product*([$p[1]$ for p in seq])
    if $seq\_score < mth$ **then**
      continue
    for p in seq **do**
      if $p[0]$ not in $ScoreMap_{multi}.keys()$ or $seq\_score > ScoreMap_{multi}[p[0]]$
      then
        $ScoreMap_{multi}[pair[0]]$ = seq_score
  NormalizeScores($ScoreMap_{single}$)
  NormalizeScores($ScoreMap_{multi}$)
  $ScoreMap_{hybrid} = \{\}$
  for id in union(set($ScoreMap_{single}.keys()$), set($ScoreMap_{multi}.keys()$))
  do
    if id not in $ScoreMap_{single}.keys()$ **then**
      $ScoreMap_{single}[id]$ = *minValue*($ScoreMap_{single}$)
    if id not in $ScoreMap_{multi}.keys()$ **then**
      $ScoreMap_{multi}[id]$ = *minValue*($ScoreMap_{multi}$)
    $ScoreMap_{hybrid}[id]$ = $ScoreMap_{single}[id]$ + $\gamma * ScoreMap_{multi}[id]$
  sorted_evi = *sortByValue*($ScoreMap_{hybrid}$)
  return sorted_evi

4. Experimental Setup

This section describes the data and setup we used for our experiments.

Split (#multi-hop)	SUPPORTS	REFUTES	NEI
Train (20,201)	80,035	29,775	35,639
Dev (1,960)	6,666	6,666	6,666
Test	6,666	6,666	6,666

Table 1: Statistics of FEVER Dataset. The number of multi-hop claims in the blind testing set is unknown.

Dataset Our experiments use a large-scale public fact verification dataset FEVER (Thorne et al., 2018), which involves retrieving multi-hop sentence-level evidence from a large text corpus before predicting a claim's verdict. The FEVER database comprises 185,455 annotated claims and 5,416,537 Wikipedia documents from the June 2017 Wikipedia dump. On average, each document contains 5 sentences. Annotators classify all claims as `SUPPORTS`, `REFUTES`, or `NOT ENOUGH INFO` based on single-hop and/or multi-hop sentence-level evidence. The dataset partition is kept the same with the FEVER Shared Task (Thorne et al., 2018) as shown in Table 1.

Evaluation Metrics As in previous work, retrieval results are compared using recall@5. Label Accuracy (LA) and FEVER score are the official evaluation metrics of the FEVER dataset. The LA metric is used to calculate the claim classification accuracy rate without considering retrieved evidence. The FEVER score checks if a complete set of golden evidence is included in the top 5 evidence retrievals in addition to the correct verdict prediction, indicating both retrieval and claim classification ability.

Implementation Details Our best dense sentence retrievers are bi-encoder models initiated from DPR-MultiData (Karpukhin et al., 2020). The negative examples are sampled using BM25 (Lin et al., 2021) and then filtered using a pre-trained attention-based sentence ranking model[1] at a threshold based on empirical criteria. In particular, samples with a similarity of over 0.999 are filtered out. The top two negative examples are then kept for training. Our best model is trained with a batch size of 512 and a max sequence length of 256. The FAISS(Johnson et al., 2017) exact inner product search index (IndexFlatIP) is used to predict retrieval results that support parallel searching in GPUs.

RoBERTa-large (Liu et al., 2019a) is trained for the sentence reranking module. Ten negative (`NOT ENOUGH INFO`) examples are sampled from the top 100 DSR retrievals for each claim. At inference time, we rerank the top 200 sentences retrieved from the last step (DSR). For the final verdict prediction, we train BEVERS's (DeHaven and Scott, 2023) claim classifier (DeBERTa-V2-XL-MNLI (He et al., 2020) + XGBoost (Chen and Guestrin, 2016)) with data constructed by pairing claims with M3's retrievals. The experiments are all conducted on a machine with 8 80GB A100 GPUs. We used Huggingface Transformers (Wolf et al., 2020) as the basis for our code.

5. Results

Following the workflow of our multi-hop retriever (M3), we report the evaluation of the five major components in M3 sequentially: (1) Single-hop Dense Sentence Retrieval, (2) Single-hop Sentence Reranking, (3) Multi-hop Dense Sentence Retrieval, (4) Multi-hop Sentence Reranking, and (5)Dynamic Hybrid Ranking.

5.1. Evidence Retrieval

A summary of the major retrieval evaluation results can be found in Table 2. We evaluate the multi-hop and the overall retrieval performance at two levels of granularity, namely the document and sentence levels. Results include non-iterative retrievers, covering sparse retrieval (BM25), dense passage retrieval (DPR), and multi-stage retrieval methods, i.e., document retrieval + sentence selection (reranking). MediaWiki API[2] is one of the most commonly used methods for document retrieval. It searches through the titles of all Wikipedia articles for entries that match the entity mentions found in the claim. Jiang et al. (2021) combine the BM25 and WikiMedia API results by going through the two ranked lists of documents alternately, skipping duplicates, and keeping the top k unique documents. Different attention-based sentence selection (reranking) methods are used, such as Enhanced Sequential Inference Model (ESIM) (Hanselowski et al., 2018), BERT (Soleimani et al., 2020), and T5 (Jiang et al., 2021).

We further compare M3 against state-of-the-art iterative retrieval approaches, including MDR, and AdMIRaL which have been introduced in Section 2. We also compare with those methods that rely on hyperlinks for multi-hop retrieval (Nie et al., 2019a; Stammbach, 2021; DeHaven and Scott, 2023). Specifically, (Stammbach, 2021) used MediaWiki API for single-hop document retrieval, while (Nie et al., 2019a) and (DeHaven and Scott, 2023) used complex combined methods, i.e., (Nie et al.,

[1] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-12-v2

[2] https://www.mediawiki.org/wiki/API:Main_page

Model Type	Model	Document-level (Rec@5) multi-hop	Overall	Sentence-level (Rec@5) multi-hop	Overall
Non-Iterative	BM25 (Lin et al., 2021)	0.252	0.714	0.385	0.614
	DPR-NQ (Karpukhin et al., 2020)	0.432	0.739	0.309	0.631
	DPR-MultiData (Karpukhin et al., 2020)	0.452	0.774	0.320	0.671
	MediaWiki API + ESIM (Hanselowski et al., 2018)	0.538	–	–	0.871
	MediaWiki API + BERT (Soleimani et al., 2020)	–	–	–	0.884
	MediaWiki API + BM25 + T5 (Jiang et al., 2021)	–	–	–	0.905
	M3-DSR$_{\text{single}}$ (ours)	0.522	0.900	0.419	0.847
	M3-DSR$_{\text{single}}$+SSR$_{\text{single}}$ (ours)	0.633	0.933	0.572	0.920
Iterative	KM + Pageview + dNSMN + sNSMN + Hyperlink (Nie et al., 2019a)	–	0.886	–	0.868
	MediaWiki API + BigBird + Hyperlink (Stammbach, 2021)	0.667	<u>0.945</u>	–	0.936
	TF-IDF + FSM + RoBERTa + Hyperlink (DeHaven and Scott, 2023)	–	–	–	**0.944**
	MDR (Xiong et al., 2021b)†	0.691	–	–	–
	AdMIRaL (Aly and Vlachos, 2022)*	<u>0.705</u>	**0.956**	–	–
	M3-full (ours)	**0.790**	**0.956**	**0.719**	<u>0.940</u>

Table 2: Retrieval performance on the FEVER dev set. DPR-NQ and DPR-MultiData indicate the DPR model trained on the NQ dataset and the DPR-MultiData dataset, respectively by (Karpukhin et al., 2020). DPR-MultiData dataset is a combination of multiple open-domain QA datasets consisting of NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), TREC (Baudis and Sedivý, 2015), and WQ (Berant et al., 2013). 'KM' = Keyword Matching. 'FSM' = Fussy String Matching. **Bold** numbers indicate best and <u>underline</u> the second-best score. Iterative models are evaluated in a two-hop process, i.e., one more hop retrieval than non-iterative models.

System	Test LA	Test FEVER
Athene (Hanselowski et al., 2018)	0.6546	0.6158
UNC NLP (Nie et al., 2019a)	0.6821	0.6421
BERT-FEVER (Soleimani et al., 2020)	0.7186	0.6966
KGAT (Liu et al., 2019b)	0.7407	0.7038
LisT5 (Jiang et al., 2021)	0.7935	0.7587
BigBird-FEVER (Stammbach, 2021)	0.7920	0.7680
ProoFVer Krishna et al. (2022)	0.7947	0.7682
BEVERS (DeHaven and Scott, 2023)	<u>0.8035</u>	**0.7786**
M3-FEVER (ours)	**0.8054**	<u>0.7743</u>

Table 3: Full system comparison for label accuracy (LA) and FEVER score on the blind FEVER test set. **Bold** numbers indicate the best and <u>underline</u> the second-best score.

2019a) used keyword matching + pageview frequency + neural network-based document reranker (dNSMN); (DeHaven and Scott, 2023) combined results from TF-IDF on titles, TF-IDF on content, and fuzzy string matching on titles queried by entities extracted from claims. Different neural network-based pairwise scoring models are used for sentence reranking, i.e., sNSMN (Nie et al., 2019a), BigBird (Stammbach, 2021), and Roberta-large (DeHaven and Scott, 2023).

As M3 supports only sentence-level retrieval, when calculating document-level recall@5, we compare the golden document IDs with those of our top 5 sentence retrievals. It's important to note that this is a harder setting for us, since on average, only 2.87 document IDs are included in our top 5 sentence retrievals. Despite this setting, M3 still achieves the highest multi-hop and overall retrieval recall for documents, and the highest multi-hop retrieval recall for sentences, only trailing (DeHaven and Scott, 2023) on retrieval recall for overall sentences.

5.2. End-to-end Fact Verification

Furthermore, we test our fact verification system M3-FEVER in an end-to-end manner on the FEVER blind test set. As shown in Table 3, M3-FEVER achieves the highest LA score and the second-best FEVER score[3].

6. Analysis

This section examines the effects of different major design decisions made in our M3 system.

[3]The FEVER Official Leaderboard: https://codalab.lisn.upsaclay.fr/competitions/7308

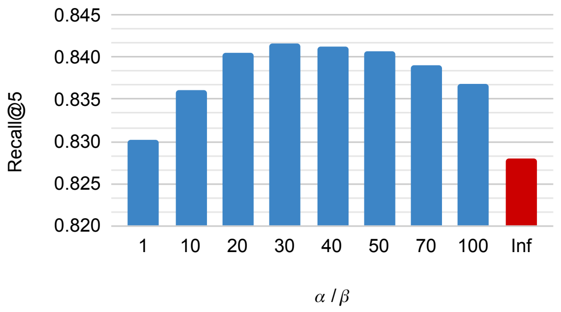

Figure 6: M3-DSR$_{\text{single}}$ top-5 retrieval recall with different ratios of multi-task learning loss weights, where α and β represent the weight of contrastive loss and claim classification loss, respectively. 'Inf' indicates only using the contrastive objective during training, i.e., single-task learning.

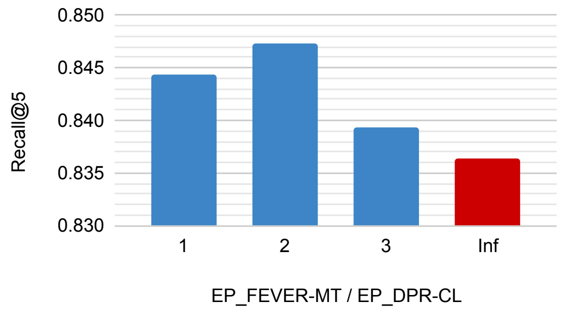

Figure 7: M3-DSR$_{\text{single}}$ top-5 retrieval recall with different ratios of mixed-objective training epochs. 'Inf' indicates that only the FEVER dataset is used for training with the multitask learning objective.

6.1. Effect of multi-task learning

We explore what ratio of multi-task learning loss weights, i.e.α/β in Equation 4, is optimal for dense sentence retrieval. Figure 6 illustrates the top-5 sentence retrieval recall of M3-DSR$_{\text{single}}$ with respect to different α/β, measured on the FEVER dev set, where α and β represent the weight of contrastive loss and claim classification loss, respectively. As is shown, when $\alpha/\beta = 30$ gives the highest retrieval recall. It outperforms the sole contrastive object learning by 1.65%. This suggests that our multi-task learning framework is effective in learning higher-quality dense sentence representations.

6.2. Effect of mixed-objective learning

We trained M3-DSR on two datasets with different objectives jointly using mixed-objective learning. Specifically, we train M3-DSR alternatively on the DPR-MultiData dataset with the contrastive objective and on FEVER with the multi-task objective. We test different ratios of training epochs for the two datasets when doing mixed-objective learning in Figure 7 and observe that when EP$_{\text{FEVER-MT}}$ / EP$_{\text{DPR-CL}}$ = 2 (i.e., training on the FEVER dataset with the multi-task object for two epochs after every epoch of training on the DPR-MultiData dataset with the contrastive object) gives the best performance. This indicates that our mixed-objective learning framework is effective at learning dense sentence representations with higher quality.

6.3. Effect of hybrid-ranking algorithm

We compare our hybrid-ranking method with two different types of merging algorithms: 1) Threshold: jointly rank multi-hop retrievals whose scores are larger than a threshold with the single-hop retrievals. 2)Scale (Stammbach, 2021; DeHaven and Scott, 2023): re-scale the multi-hop retrievals by a factor before jointly ranking them together with the single-hop retrievals. Table 4 demonstrates that our hybrid-ranking algorithm outperforms the baseline algorithms by a large margin.

Method	Recall@5
Threshold	0.925
Scale	0.931
Hybrid Ranking	**0.940**

Table 4: Ablation of the hybrid ranking algorithm over the FEVER's dev set. All hyperparameters are tuned through grid search over our best SRR$_{\text{multi}}$'s results.

6.4. Effect of negative sampling

Due to the difficulty of exhaustively annotating all positive examples given a query, false negatives are common in large-scale retrieval datasets. Figure 8 demonstrated false negative examples in the FEVER dataset. When using traditional sampling methods such as BM25 to sample negative examples for contrastive learning, we find it difficult to avoid false negatives. By applying an empirical threshold to an off-the-shelf attention-based ranking model, we can eliminate more false negatives from training data, thereby further improving M3-DSR$_{\text{single}}$'s recall by 5.6%.

7. Conclusion

In this paper, we introduce M3, an advanced recursive multi-hop dense sentence retrieval system designed for fact verification. M3 achieves top-tier performance in multi-hop retrieval on the FEVER dataset. We propose a novel method for learning dense sentence representations, which is based

Claim: Romelu Lukaku plays in the Premier League for Everton.
Single-hop evidence annotations: 1. (Title: Romelu Lukaku) Romelu Menama Lukaku (born 13 May 1993) is a Belgian professional footballer who plays as a striker for Premier League club Everton and the Belgium national team. 2. (Title: Romelu Lukaku) He did not appear regularly in his first season there, and spent the following two seasons on loan at West Bromwich Albion and Everton respectively, signing permanently for the latter for a club record # 28 million in 2014.
Top-2 sampled negatives by BM25: 1. (Title: Lukaku) *Romelu Lukaku (born 1993), Belgian footballer, who currently plays for Everton.* 2. (Title: Roger Lukaku) He is the father of footballers Romelu Lukaku and Jordan Lukaku.
Verdict: Supported

Figure 8: An example of a false negative sampled by BM25 from the FEVER is highlighted in red.

on multi-task learning and mixed-objective learning. This approach addresses challenges faced by current dense retrieval methods that rely on contrastive learning. Furthermore, we present an efficient heuristic hybrid ranking algorithm that combines single-hop and multi-hop sentence evidence, resulting in significant improvements over previous methods. Lastly, we develop an end-to-end multi-hop fact verification system built upon M3, which also attains state-of-the-art performance on the FEVER dataset.

8. Acknowledgements

Our sincere thanks go out to the anonymous reviewers who took the time to provide constructive comments.

9. Ethics Statement

We expect our retrieval system to be utilized in fact-checking applications. It's important to clarify that our system doesn't assess the accuracy of real-world statements; it solely relies on Wikipedia as its source of evidence, as our entire testing environment is limited to this dataset. While Wikipedia is a valuable collaborative resource, it's not immune to errors and inaccuracies, just like any other encyclopedia or knowledge base. Therefore, we advise users against using our retrieval system to make definitive claims about the accuracy of the statements being verified. In other words, it should not be employed as a tool to determine the absolute truth of claims – in other words, avoid using it to declare statements as true or false.

10. References

Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Cross-domain modeling of sentence-level evidence for document retrieval. 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 3490–3496, Hong Kong, China. Association for Computational Linguistics.

Rami Aly and Andreas Vlachos. 2022. Natural logic-guided autoregressive multi-hop document retrieval for fact verification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6123–6135, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2019. Learning to retrieve reasoning paths over wikipedia graph for question answering. *ArXiv*, abs/1911.10470.

Yang Bai, Anthony Colas, and Daisy Zhe Wang. 2023. Mythqa: Query-based large-scale check-worthy claim detection through multi-answer open-domain question answering. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

P. Baudis and J. Sedivý. 2015. Modeling of the question answering task in the yodaqa system. In *CLEF*.

Jonathan Berant, Andrew K. Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *EMNLP*.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Mitchell DeHaven and Stephen Scott. 2023. BEVERS: A general, simple, and performant framework for automatic fact verification. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, pages 58–65, Dubrovnik, Croatia. Association for Computational Linguistics.

Martin Fajcik, Petr Motlicek, and Pavel Smrz. 2023. Claim-dissector: An interpretable fact-checking system with joint re-ranking and veracity prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10184–10205, Toronto, Canada. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2:1735–1742.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Kelvin Jiang, Ronak Pradeep, Jimmy J. Lin, and David R. Cheriton. 2021. Exploring listwise evidence reasoning with t5 for fact verification. In *Annual Meeting of the Association for Computational Linguistics*.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Kumar Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. In *Findings*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7:535–547.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Conference on Empirical Methods in Natural Language Processing*.

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*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022. Proofver: Natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics*, 10:1013–1030.

T. Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, D. Epstein, Illia Polosukhin, J. Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc V. Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Xin-Yi Li, Wei-Jun Lei, and Yu-Bin Yang. 2023. From easy to hard: Two-stage selector and reader for multi-hop question answering. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2356–2362, New York, NY, USA. Association for Computing Machinery.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2019b. Fine-grained fact verification with kernel graph attention network. In *Annual Meeting of the Association for Computational Linguistics*.

Xueguang Ma, Kai Sun, Ronak Pradeep, and Jimmy J. Lin. 2021. A replication study of dense passage retriever. *ArXiv*, abs/2104.05740.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019a. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Yixin Nie, Songhe Wang, and Mohit Bansal. 2019b. Revealing the importance of semantic retrieval for machine reading at scale. *ArXiv*, abs/1909.08041.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. In *North American Chapter of the Association for Computational Linguistics*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*.

Juan Enrique Ramos. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2825–2835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. In *Text Retrieval Conference*.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. Bert for evidence retrieval and claim verification. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II*, page 359–366, Berlin, Heidelberg. Springer-Verlag.

Dominik Stammbach. 2021. Evidence selection as a token-level prediction task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 14–20, Dominican Republic. Association for Computational Linguistics.

Shyam Subramanian and Kyumin Lee. 2020. Hierarchical Evidence Set Modeling for automated fact extraction and verification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7798–7809, Online. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. 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 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Bohong Wu, Zhuosheng Zhang, Jinyuan Wang, and Hai Zhao. 2021. Sentence-aware contrastive learning for open-domain passage retrieval. In *Annual Meeting of the Association for Computational Linguistics*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021a. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Wenhan Xiong, Xiang Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Scott Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oguz. 2021b. Answering complex open-domain questions with multi-hop dense retrieval. In *International Conference on Learning Representations*.

Peilin Yang, Hui Fang, and Jimmy J. Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. UCL machine reading group: Four factor framework for fact finding (HexaF). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 97–102, Brussels, Belgium. Association for Computational Linguistics.

Hang Zhang, Yeyun Gong, Yelong Shen, Jiancheng Lv1, Nan Duan, and Weizhu Chen. 2022. Adversarial retriever-ranker for dense text retrieval. In *International Conference on Learning Representations*.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, M. Zhou, Jiahai Wang, and Jian Yin. 2019. Reasoning over semantic-level graph for fact checking. In *Annual Meeting of the Association for Computational Linguistics*.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy. Association for Computational Linguistics.