UF Herbert Wertheim College of Engineering
UNIVERSITY of FLORIDA

DSR @ UF
Data Science Research

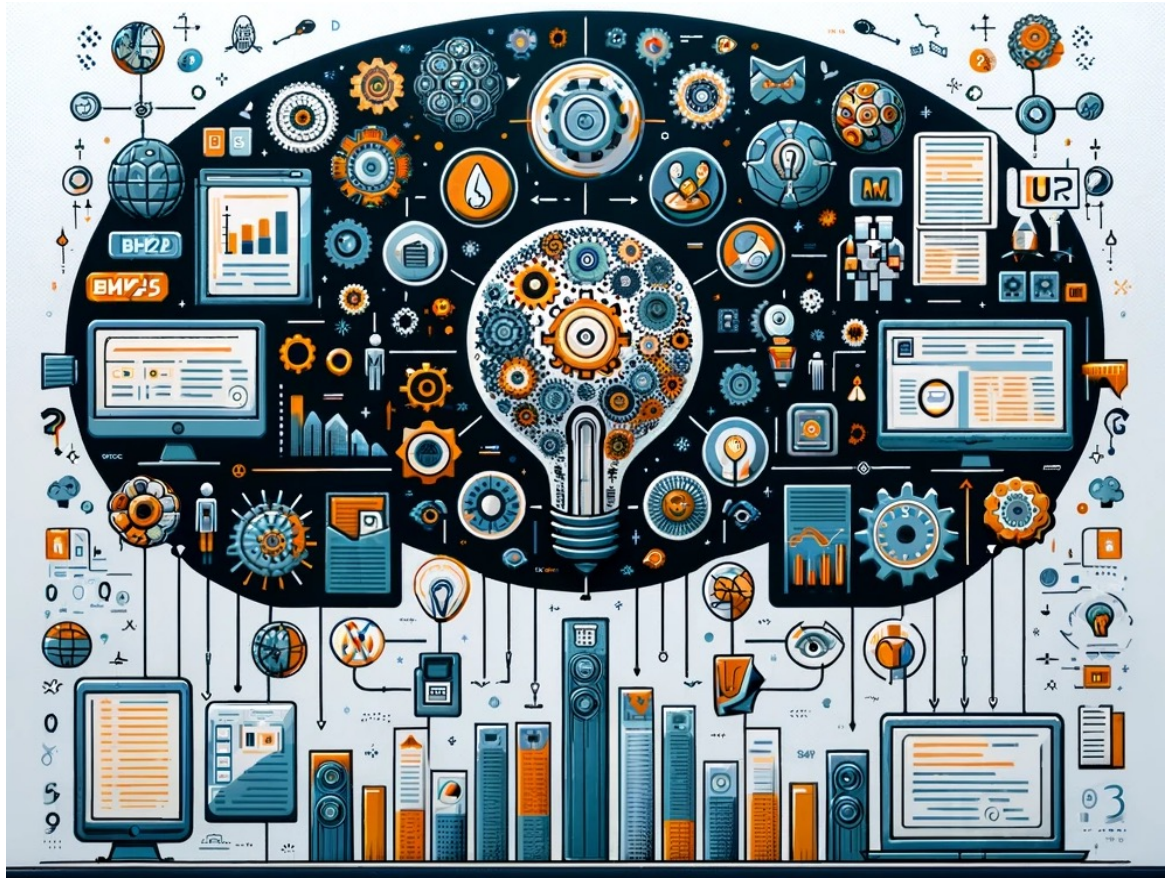
M3: A Multi-Task Mixed-Objective Learning Framework for Open-Domain Multi-Hop Dense Sentence Retrieval

LREC-CONLING 2024

Yang Bai, Anthony Colas, Christan Grant, Daisy Zhe Wang
University of Florida

Agenda

1. Introduction
2. Methods
3. Experiments
4. Ablation Studies
5. Conclusion



Background: Large-Scale Fact Verification

- Large-scale fact verification is a challenging task where ***single-hop or multi-hop sentence-level evidence*** needs to be retrieved from ***a large pool of documents*** to verify human-generated ***claims***.

- A three-stage approach is commonly used to solve the problem.

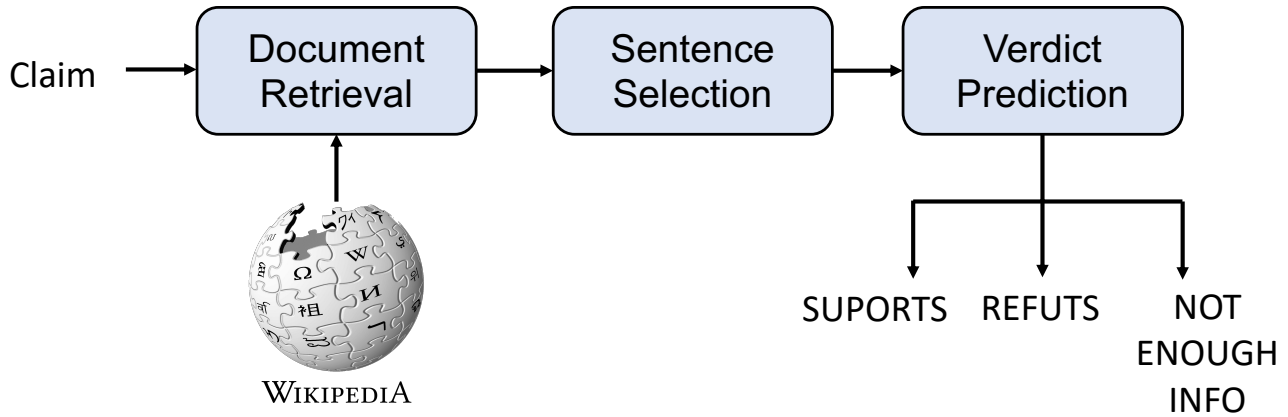


Figure 1: Canonical thee-stage fact verification framework.

Background: FEVER Dataset

- The ***FEVER (Fact Extraction and VERification)*** dataset stands as one of the most ***challenging and renowned*** benchmark datasets for ***large-scale fact verification***.

- The FEVER dataset demonstrates its prominence and difficulty through the ***extensive engagement*** it has received from the research community. Since it was published in ***2018***, it has received ***over 13K citations***.

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

Current State-of-the-art Systems on FEVER dataset

- **BERVERS**: ***TF-IDF + Fuzzy String Matching (FSM)*** for document retrieval. It featured by using ***Hyperlinks*** for multi-hop evidence retrieval.
  - ***Challenges***: ***Relying on hyperlinks*** prevents the model from being generative since it requires data that contains hyperlinks.

- **AdMIRaL**: ***BM25*** for document retrieval. It featured by using an ***autoregressive model*** for iterative document retrieval.
  - ***Challenges***: The ***autoregressive model must be retrained for each new dataset*** since it predicts document IDs, which limits its ability to be applied more broadly.

Research Question

- How to ***enhance existing dense retrieval methods*** to ***boost the recall of evidence retrieval***, thereby ***improving the overall accuracy of the end-to-end fact-verification*** process on the FEVER dataset?

Challenges

1. Train with ***the passage level data*** may learn flawed representations because of ***internal conflicts***.
   a. ***passages may contain multiple semantically different even opposite sentences***
   b. may cause conflicts in contrastive learning, especially when large batch sizes are used for ***in-batch negative sampling.***
2. ***Soly rely on contrastive objectives*** may hinder optimal representation learning and recall.
3. ***Training over datasets with different objective coherently in mixed objective learning setting can be challenging***.

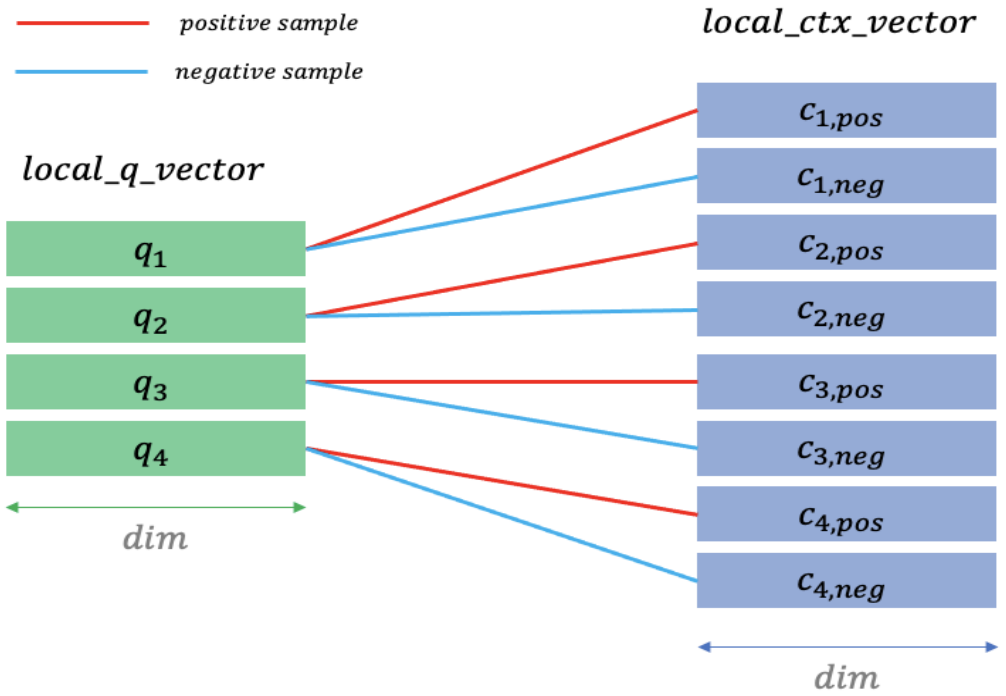


Fig3. In-batch negative sampling [2].

[2] https://user-images.githubusercontent.com/42966248/111322063-dec64f00-86ab-11eb-977c-a83f4d5bb98c.png

Contributions

- We present an advanced recursive ***multi-hop dense sentence retrieval system (M3)*** based on a novel dense sentence representation learning method, which achieves competitive 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*** and ***DPR*** on sentence-level retrieval.
- We introduce a ***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.

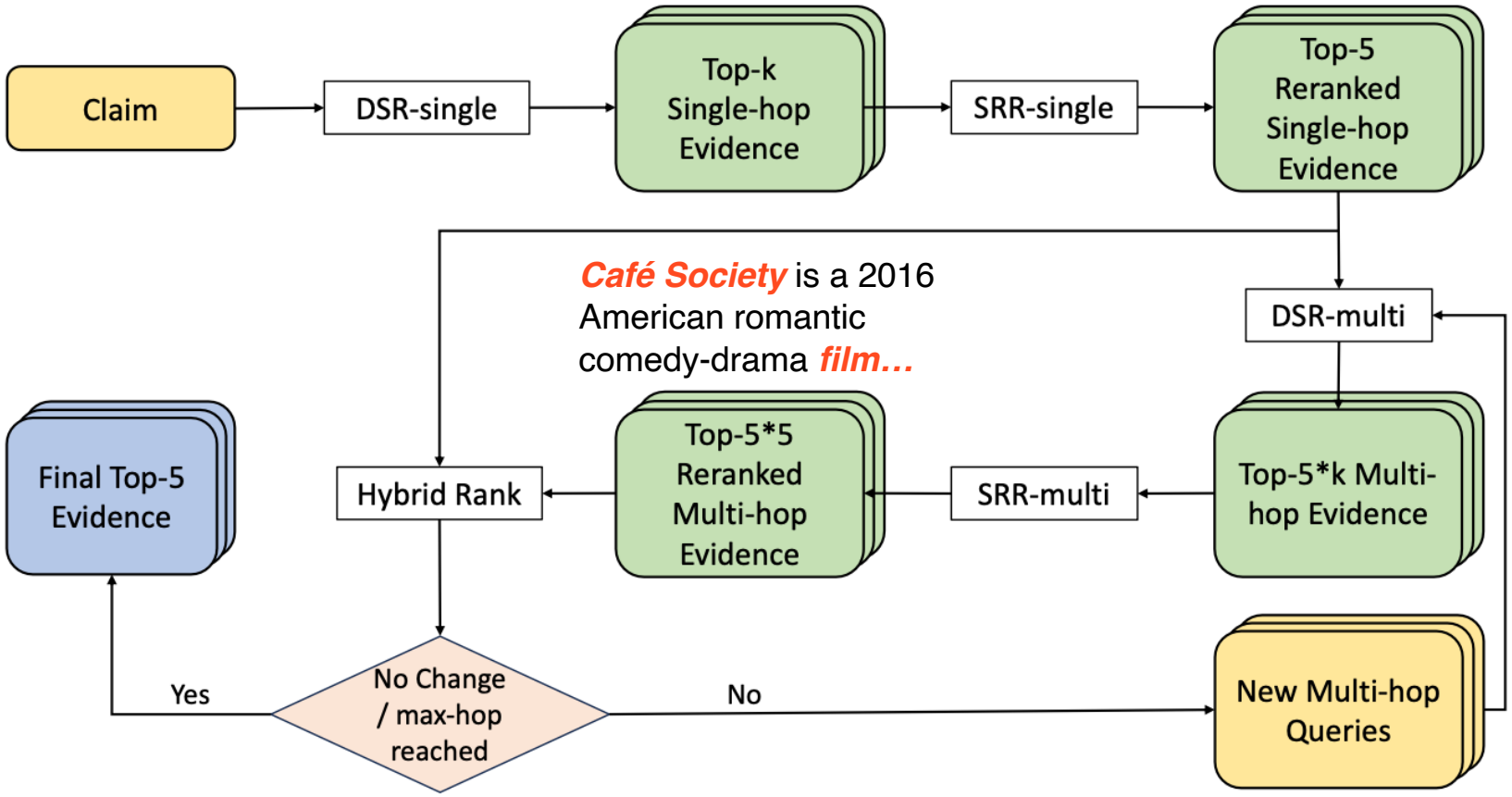
System Overview

Claim: ***Sheryl Lee*** has yet to appear in a ***film*** as of ***2016***.

1. ***Sheryl Lee***-In ***2016***, she appeared in ***Café Society***…

2. ***Café Society*** is a 2016 American romantic comedy-drama ***film…***

3. …



Café Society is a 2016 American romantic comedy-drama ***film…***

Sheryl Lee -- In ***2016***, she appeared in ***Café Society***…

Claim: ***Sheryl Lee*** has yet to appear in a ***film*** as of ***2016***.

⊕

Sheryl Lee–In ***2016***, she appeared in ***Café Society***…

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.

Multi-task Learning

- **Contrastive Learning Objective**

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

- **Classification Objective**

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

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

- **Multi-Task Learning Objective**

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

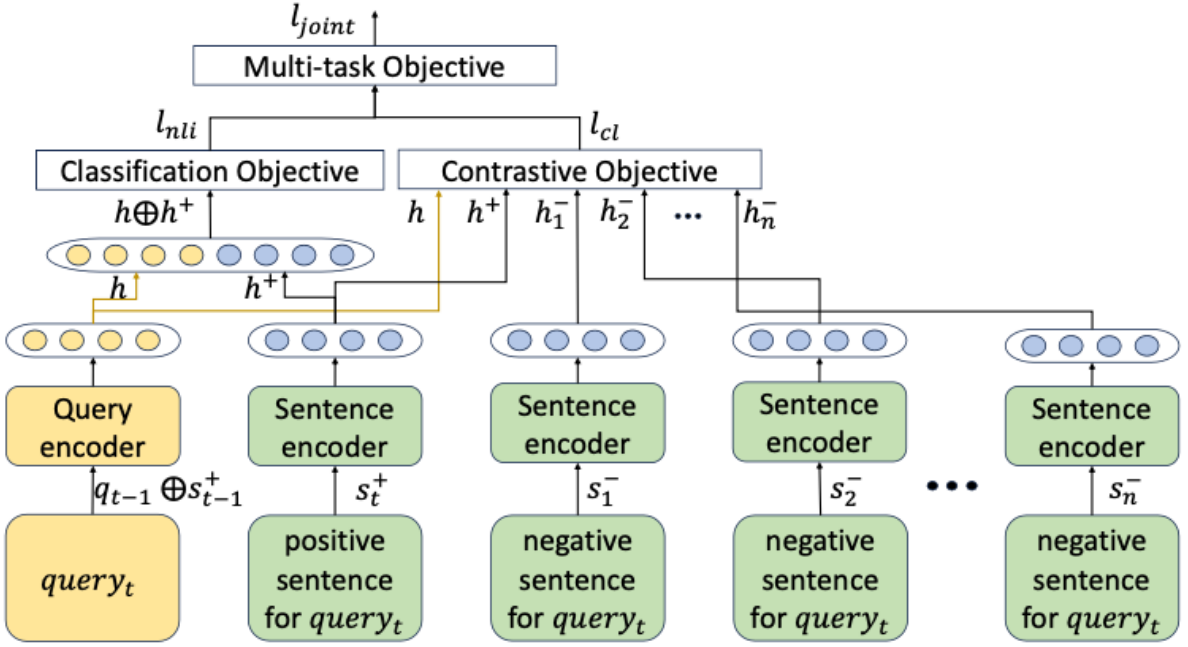


Figure 4: M3-DSR multi-task learning framework. When $t = 1$ (i.e., first-hop), the input of the query encoder is the original claim c.

Mixed-Objective Learning

- To maximize the utility of diverse datasets for learning dense representations, ***we've developed a framework that integrates multiple learning objectives***

- It enables us to ***consistently*** train a model on ***various datasets***, ***each with its own objective function***, within a ***single training session*** by adjusting the ***frequency*** of each dataset's use.

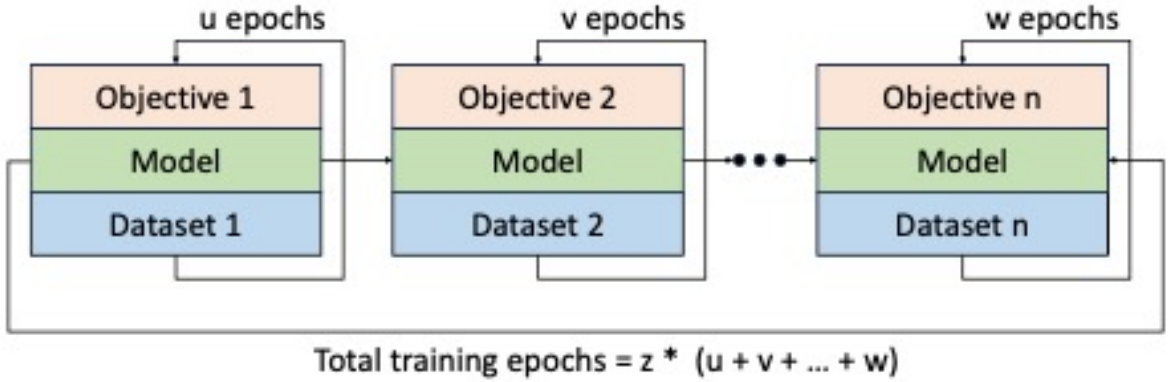


Figure 5: M3-DSR mixed-objective learning framework. The same model is trained with different dataset-objective combinations sequentially.

Methods: Hybrid Ranking

- To optimize retrieval recall ***in FEVER, where not all claims require multi-hop evidence***, we introduce a hybrid ranking algorithm to efficiently ***combine single-hop and multi-hop retrievals***.

- We first ***scale*** the retrieval score for each step of multi-hop retrieval through ***production*** to make sure that each episode of evidence is ***proportional to the other***.

- Then the single-hop and the multi-hop evidence scores are combined through a ***weighted sum***. The weight is a ***hyperparameter***.

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[p[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] +
                γ * ScoreMap_multi[id]
    sorted_evi = sortByValue(ScoreMap_hybrid)
    return sorted_evi
```

Results: Retrieval

Model Type	Model	Document-level (Rec@5) multi-hop	Document-level (Rec@5) Overall	Sentence-level (Rec@5) multi-hop	Sentence-level (Rec@5) 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.

BEVERS

Results: FEVER Leaderboard

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)	0.8035	**0.7786**
M3-FEVER (ours)	**0.8054**	0.7743

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

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.

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

- When ***α/β = 30 gives the highest retrieval recall***.
- This suggests that ***our multi-task learning framework is effective*** in learning higher-quality dense sentence representations.

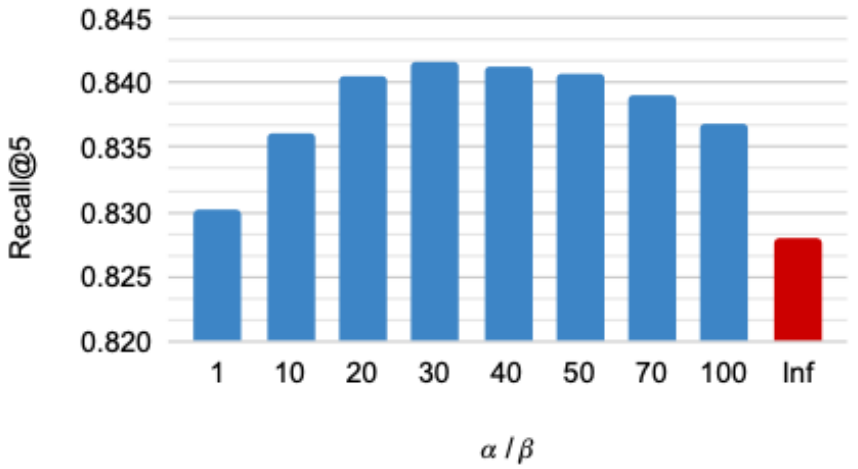


Figure 6: M3-DSR$_{\text{single}}$ top-5 retrieval recall with different ratios of multi-task learning loss weights, where α and β represent the weight of contrastive loss and claim classification loss, respectively. 'Inf' indicates only using the contrastive objective during training, i.e., single-task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

Effect of mixed-objective learning

- In our study of mixed-objective learning, we found that ***allocating twice as many training epochs to the FEVER dataset as to the DPR-MultiData dataset results in the best retrieval recall*** for our dense retriever.
- This indicates that ***our mixed-objective learning framework is effective*** at learning dense sentence representations with higher quality.

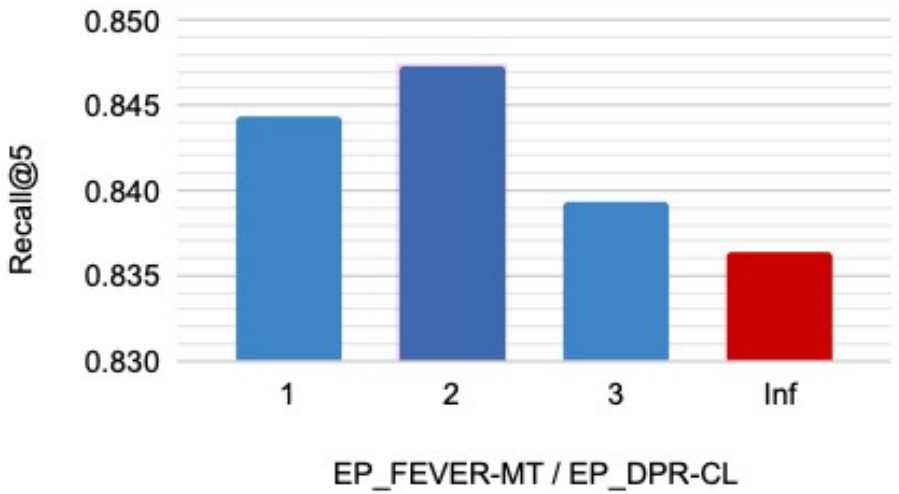


Figure 7: M3-DSR$_{\text{single}}$ top-5 retrieval recall with different ratios of mixed-objective training epochs. 'Inf' indicates that only the FEVER dataset is used for training with the multitask learning objective.

Effect of hybrid-ranking algorithm

- We compare our hybrid-ranking method with two different types of merging algorithms:
  - 1) ***Threshold***: jointly ranks multi-hop retrievals whose scores are larger than a threshold with the single-hop retrievals.
  - 2)***Scale (BEVERS, etc)***: re-scales 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***.

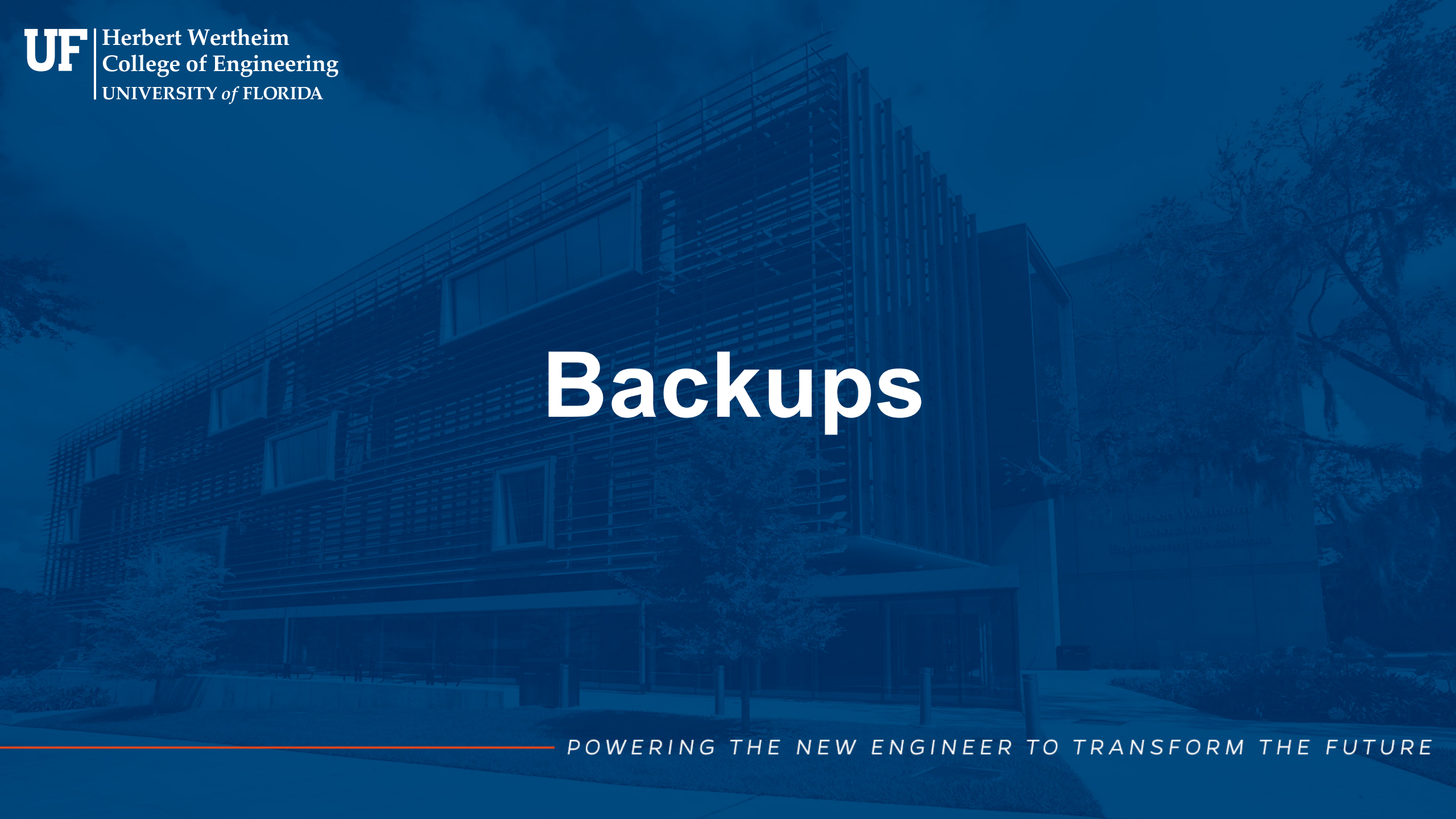
Method	Recall@5
Threshold	0.925
Scale	0.931
Hybrid Ranking	**0.940**

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

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 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 ***a heuristic hybrid ranking algorithm that combines single-hop and multi-hop sentence evidence***, resulting in substantial 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.


UF
Herbert Wertheim
College of Engineering
UNIVERSITY of FLORIDA
POWERING THE NEW ENGINEER TO TRANSFORM THE FUTURE


UF
Herbert Wertheim
College of Engineering
UNIVERSITY of FLORIDA
Backups
POWERING THE NEW ENGINEER TO TRANSFORM THE FUTURE

Information Retrieval (IR)

1. IR is the process of ***searching*** for specific information, typically ***text documents***, within ***vast digital collections*** to satisfy a particular ***query***.

2. This capability ***enhances the scalability*** of NLP tasks, enabling the transition from Machine Reading Comprehension (MRC) ***to Open-domain QA*** and evolving Stance Classification into ***Automatic Fact Verification***, etc.

3. Additionally, ***Retrieval-Augmented Generation (RAG)*** combines IR with generative LLMs to improve ***context understanding***, increase ***response accuracy***, and ***reduce 'hallucinations'*** in generated content.

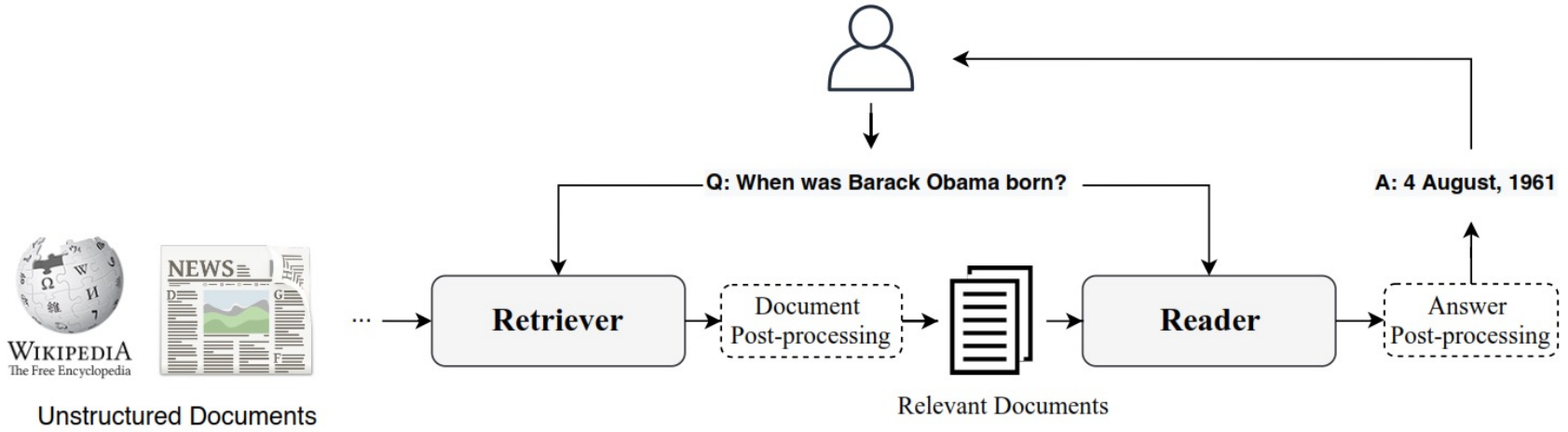


Fig.1 An illustration of "Retriever-Reader" architecture of an Open-domain QA system [1].

[1]Zhu, Fengbin et al. "Retrieving and Reading: A Comprehensive Survey on Open-domain Question Answering." *ArXiv* abs/2101.00774 (2021): n. pag.

Machine Learning (ML)-based IR

- Traditional retrieval models like TF-IDF and BM25 rely on ***exact term matching***, often ***overlooking the deeper semantic meaning*** within questions and documents.

- Machine learning approaches employ ***neural networks*** to represent the ***semantic meaning*** of questions and documents in a ***dense vector space***. For retrieval, they leverage techniques like ***approximate nearest neighbor (ANN),*** offering a significant improvement over the traditional approaches.

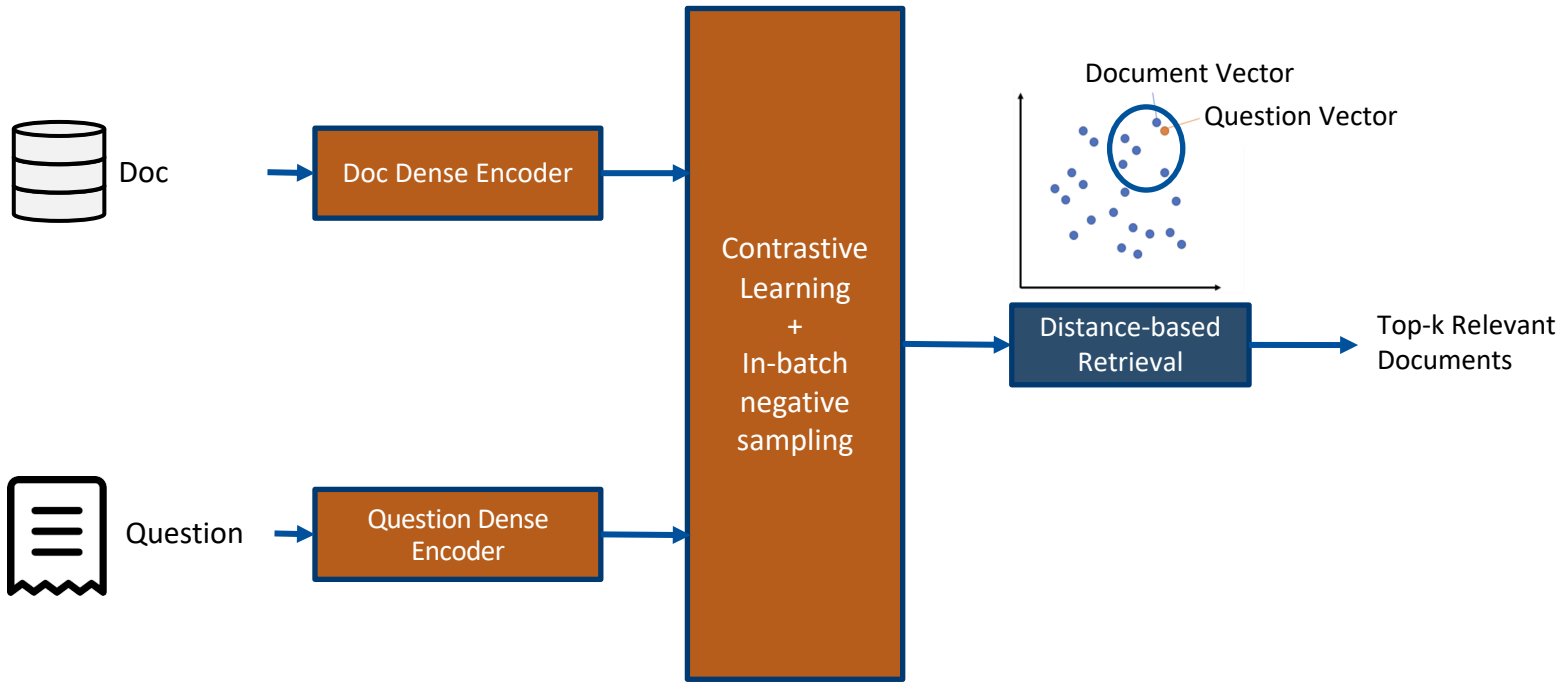


Fig.2 An illustration of a bi-encoder architecture for learning dense representations to improve retrieval

Contrastive Learning (CL)

CL adjusts the model so that ***representations of similar data points are encoded closer*** together in the ***embedding space***, while representations of ***dissimilar data points are farther*** apart.

$$\mathcal{D} = \{\langle q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^- \rangle\}_{i=1}^m$$

eq1. Data example representation

$$L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) = -\log \frac{e^{\text{sim}(q_i, p_i^+)}}{e^{\text{sim}(q_i, p_i^+)} + \sum_{j=1}^n e^{\text{sim}(q_i, p_{i,j}^-)}}$$

eq2. NLL loss function

$$\text{sim}(q, p) = E_Q(q)^\intercal E_P(p)$$

eq3. Similarity function

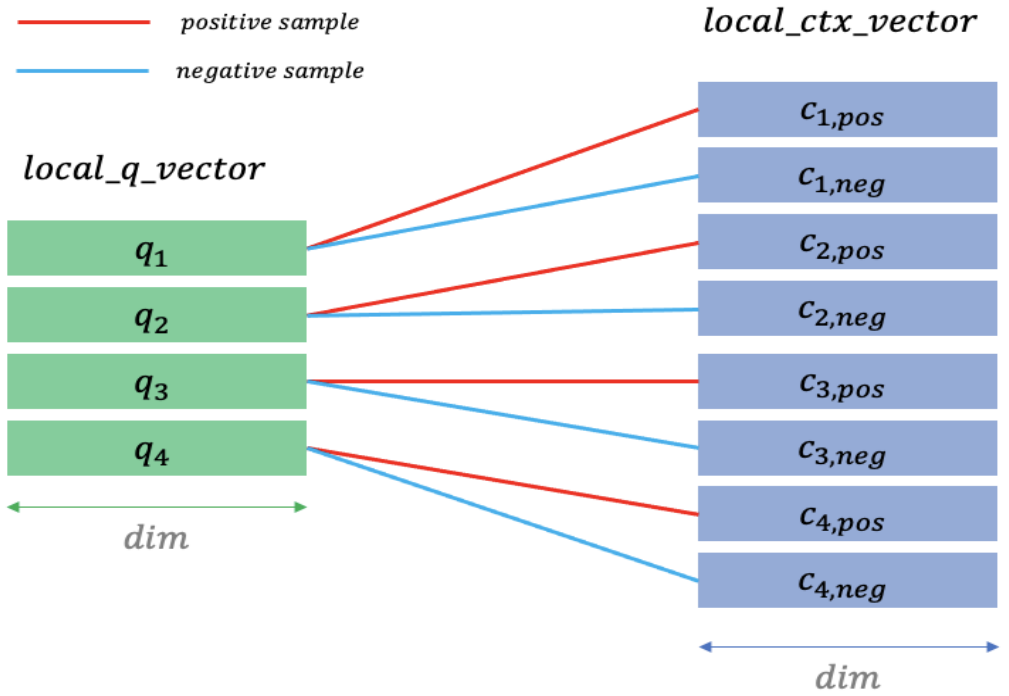


Fig3. In-batch negative sampling [2].

[2] https://user-images.githubusercontent.com/42966248/111322063-dec64f00-86ab-11eb-977c-a83f4d5bb98c.png

Implementation Details (1)

- The negative examples are ***sampled using BM25*** and then ***filtered using a pre-trained attention-based sentence ranking model*** at a threshold based on empirical threshold.
- The top ***two negative examples*** are kept for each claim for CL.
- Our best model is trained with a ***batch size of 512*** and a ***max sequence length of 256***.
- We use the ***FAISS-IndexFlatIP to organize the encoded corpus*** for ANN retrieval. It supports ***parallel retrieval in GPUs***.
- The experiments are all conducted on a machine ***with 8 80GB A100 GPUs***. We used ***Huggingface Transformers*** as the basis for our code.

Implementation Details (2)

- ***RoBERTa-large*** 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 pre- diction, we train claim classifier (***DeBERTa-V2-XL-MNLI + XGBoost***) with data constructed by pairing claims with M3's retrievals.

Analysis: 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.
- 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 's recall by ***5.6%***.

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.