Retrieve-and-Sample: Document-level Event Argument Extraction via Hybrid Retrieval Augmentation

Yubing Ren

Co-authors: Yanan Cao∗, Ping Guo, Fang Fang, Wei Ma∗, Zheng Lin
Institute of Information Engineering, CAS

Outline

- Background
- Our work
- Experience sharing

Document-level Event Argument Extraction

- Transforming the large amounts of unstructured text on the Internet into structured event knowledge is a critical, yet unsolved goal of NLP, especially when addressing document-level text.

- Document-level EAE is the process of extracting informative event kernels from a document, which benefits many downstream applications, e.g., IR, QA, and event graph reasoning.

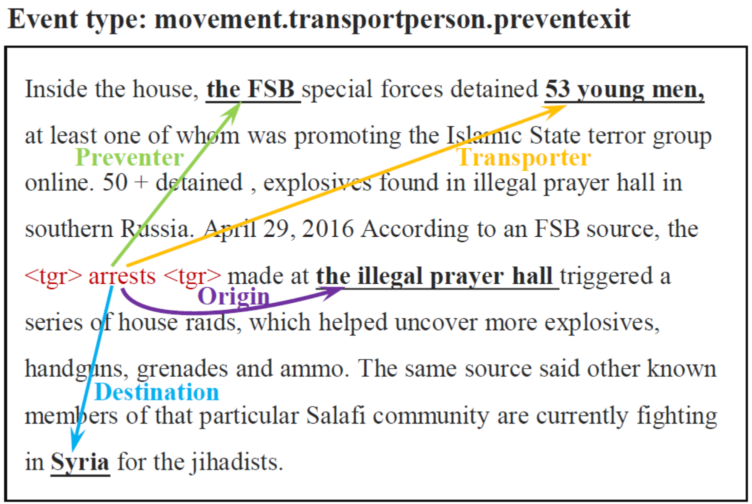


Figure 1: An illustration of document-level EAE task. Trigger words are included in special tokens <tgr>. Underlined words denote arguments, and arcs denote roles.

Retrieval augmentation in NLP

- Retrieval-augmented methods have recently been successfully applied to NLP tasks, e.g., DRG, MT and IE, which allows models to:
  - explicitly acquire prior external knowledge in a non-parametric manner, leading to great flexibility.
  - regard the retrieved reference instances as cues to generate text and learn by analogy.
- These retrieval-augmented methods use similarity-based retrieval, which bases on a simple hypothesis:
  - the more $\boldsymbol{x_r}$ (retrieved demonstration) resembles $\boldsymbol{x}$ (original input), the more likely $\boldsymbol{y_r}$ (demonstration label) resembles $\boldsymbol{y}$ (input label), so it will help the generation.
  - Intuitively, similar input results in similar output for most tasks, e.g., in language modeling, "*Dickens is the author of* " and "*Dickens wrote*" will have essentially the same distribution over the next word.

However

- In document-level EAE, $\boldsymbol{x_r}$ resembles $\boldsymbol{x}$ cannot guarantee the equivalent distribution of $\boldsymbol{y_r}$ and $\boldsymbol{y}$ in label space:
  - only a few words are event arguments, while other distracting context can mislead similarity-based retrieval and cause demonstration label $\boldsymbol{y_r}$ deviate from input label $\boldsymbol{y}$.
  - predict argument entity & the correspondence between arguments and roles, which makes it challenging to find a demonstration with an identical event label to the original doc.

- This raises an interesting question:
  - since document-level EAE doesn't satisfy the hypothesis of similarity-based retrieval, how do we design the retrieval strategy for document-level EAE?

Outline

- Background
- Our work
- Experience sharing

Explore various retrieval settings for DocEAE.

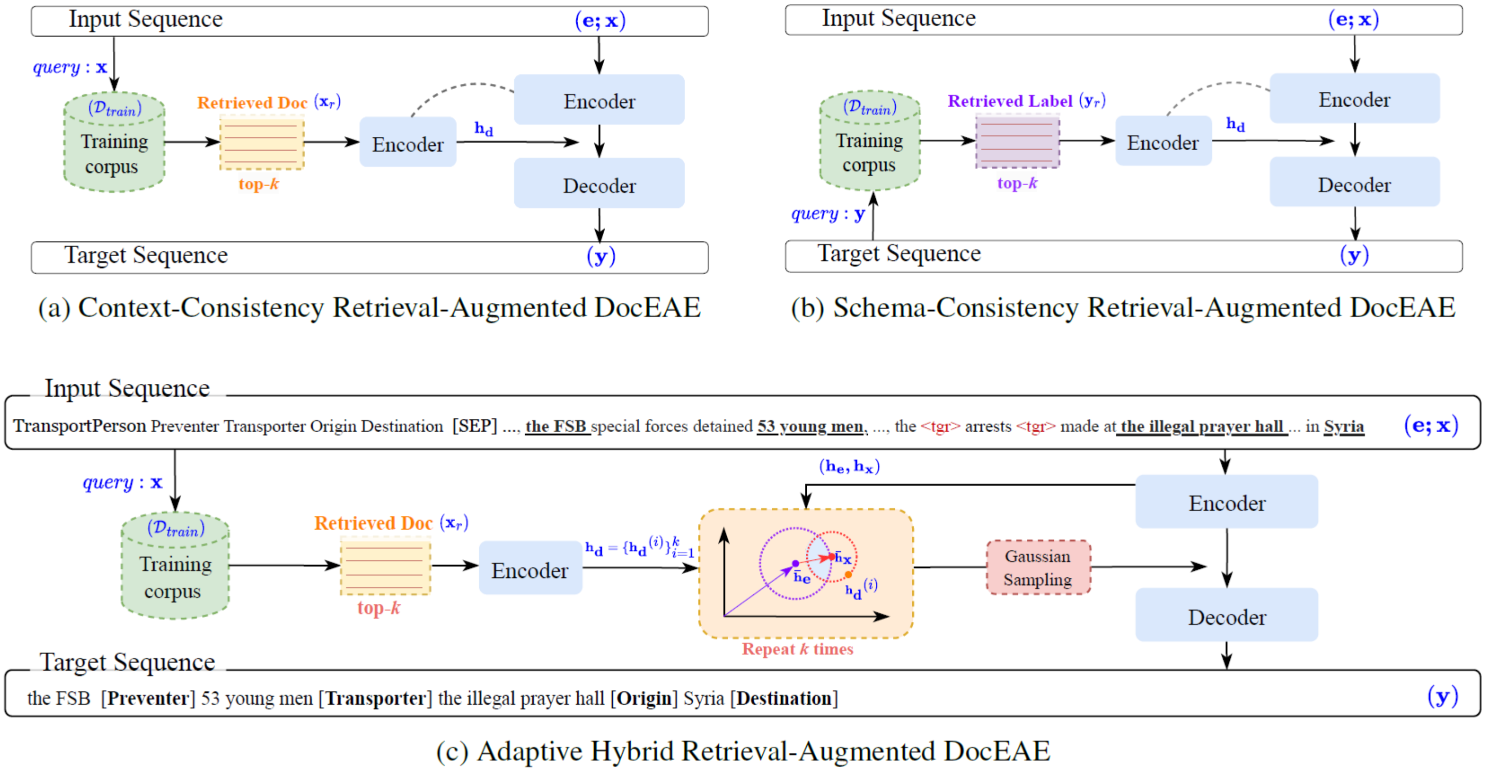
I. If similar documents cannot guarantee the same distribution of event labels, does it make sense to pursue $\boldsymbol{x_r}$ to be similar to $\boldsymbol{x}$ in retrieval process? →**Setting 1**

II. Since the essence of the above hypothesis is to pursue $\boldsymbol{y_r}$ resembles $\boldsymbol{y}$, why don't we directly retrieve $\boldsymbol{y_r}$ similar with $\boldsymbol{y}$ as the reference? → **Setting 2**

III. We want a demonstration that has equal distribution with original document in both input and label space, how should we do? → **Setting 3**

- Intuitively, it is impossible to retrieve the ideal demonstration in discrete space, so we sample a cluster of pseudo demonstrations in continuous space instead.

Overall Structure



(a) Context-Consistency Retrieval-Augmented DocEAE

(b) Schema-Consistency Retrieval-Augmented DocEAE

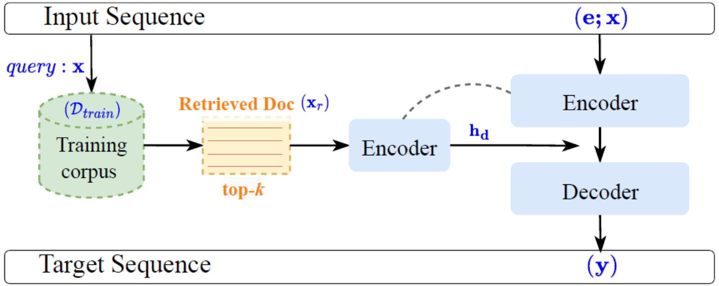
(c) Adaptive Hybrid Retrieval-Augmented DocEAE

Figure 2: An illustration of our proposed three retrieval-augmented DocEAE.

Backbone: T5.

Setting 1: Context-Consistency Retrieval

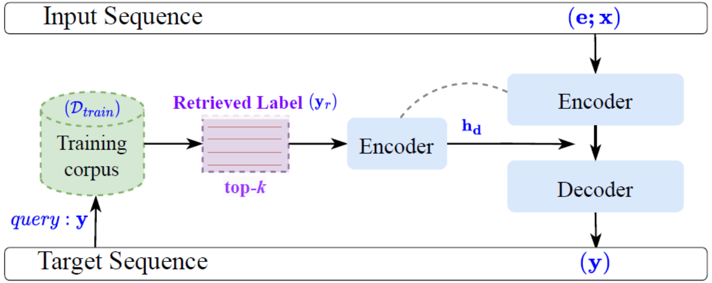
- Why?
  - Since similar documents cannot guarantee the same distribution of event labels, Setting 1 aims to answer whether it makes sense to pursue $\boldsymbol{x_r}$ to be similar to $\boldsymbol{x}$ in the retrieval process.
- How?
  - Given a query document $\boldsymbol{x}$, we retrieve the instance document $\boldsymbol{x_r}$ from the training corpus $\boldsymbol{D_{train}}$ that is the top-k relevant to the original input document, as discrete demonstrations $\boldsymbol{d}$.
  - For retrieval, we use S-BERT to retrieve semantically similar documents $\boldsymbol{x_r} \in \boldsymbol{D_{train}}$.



(a) Context-Consistency Retrieval-Augmented DocEAE

Setting 2: Schema-Consistency Retrieval

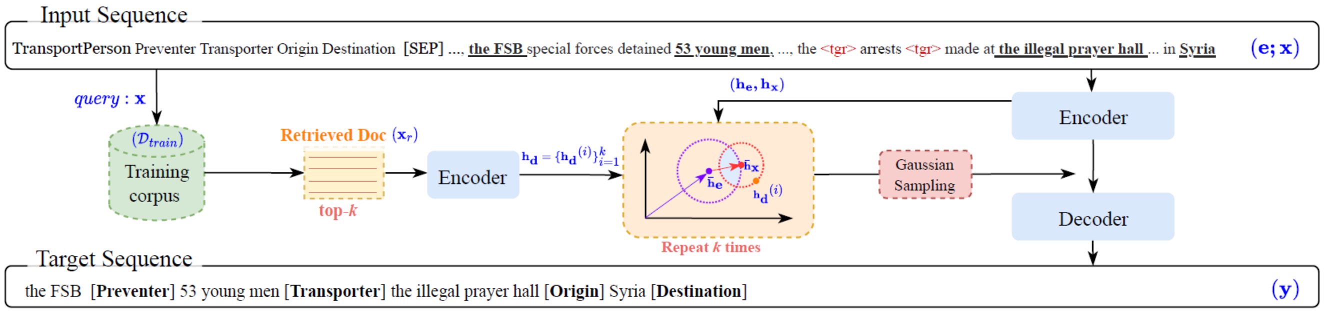
- Why?
  - To explore whether conditioning on the label space contributes to performance gains, setting 2 satisfies event schema consistency and aims to alleviate the difficulty of learning the complex event pattern of $\boldsymbol{y}$.
- How?
  - Given the event label $\boldsymbol{y}$ of input as query, we retrieve (also via S-BERT) the instance label $\boldsymbol{y_r}$ that is the top-k relevant to the input label from the training corpus $\boldsymbol{D_{train}}$.
  - During the inference, the query is the event schema e of test sample.



(b) Schema-Consistency Retrieval-Augmented DocEAE

Setting 3: Adaptive Hybrid Retrieval

- Why?
  - To find the ideal demonstration that has equal distribution with original doc in both input and label space as depth cues to guide the model and improve its analogical capability.
- How? Retrieve and Sample
  - Given an document $\boldsymbol{x}$, we first retrieve top-k helpful documents from $\boldsymbol{D}_{\boldsymbol{train}}$. Conditioning on k discrete demonstrations, we adaptively determine k event semantic regions in continuous space for each training instance. Then we sample k pseudo demonstrations from k event semantic regions.



(c) Adaptive Hybrid Retrieval-Augmented DocEAE

Setting 3： Adaptive Hybrid Retrieval

- Event Semantic Region
  - We treat points in the event semantic region as the critical states of event-semantic equivalence.
  - We first determine the adjacent region of doc and event schema by setting their adjacent radii (the orange circle and purple circle in Fig 3).
  - Then define the intersection of their adjacent regions as an event semantic region (the light blue region in Fig 3), which describes accurate variants in consistency with original context and event semantic meaning.
  - Here we have k discrete demonstration embeddings $\boldsymbol{h_d}$ for k adjacent radii r, which determines k event semantic regions.
- For each event semantic region, we perform the following Gaussian sampling.

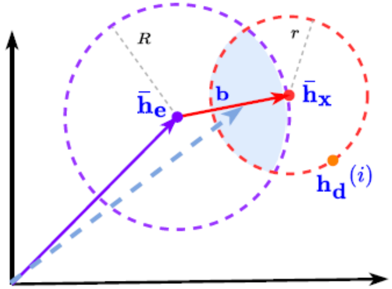


Figure 3: The geometric diagram of the proposed Gaussian sampling. $\boldsymbol{h_e}$, $\boldsymbol{h_x}$ and $\boldsymbol{h_d}$ are the representations of event schema, document and discrete demonstrations. We sample pseudo demonstrations from event semantic region (the light blue intersection).

Setting 3: Adaptive Hybrid Retrieval

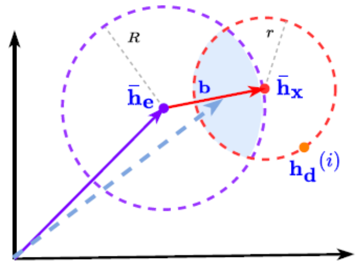
- Gaussian Sampling
  - we construct a novel sample $\boldsymbol{v} = \boldsymbol{h_e} + \omega \odot \boldsymbol{b}$ from light blue region as a pseudo demonstration, the bias vector $\boldsymbol{b} = \boldsymbol{h_x} - \boldsymbol{h_e}$.
  - The goal of sampling strategy turns into find a set of scale vectors, i.e.$\omega \in \{\omega^{(1)}, \omega^{(2)}, \ldots \omega^{(k)}\}, \omega^{(i)} \in (1 - \frac{r^{(i)}}{R}, 1)$. Intuitively, we can assume that ω follows a distribution with Gaussian forms, formally:

$$\omega^{(i)} \sim \mathcal{N}\left(\frac{1 - \frac{r^{(i)}}{R} + 1}{2}, \mathrm{diag}(\mathcal{W}_r^2)\right)$$

  - Since sampling is a non-differentiable operation that truncates the gradient, so here we use a reparametrization trick:

$$\omega^{(i)} = \mu + \epsilon \cdot \sigma, \text{ where } \epsilon \sim \mathcal{N}(0,1), \mu = \frac{1 - \frac{r^{(i)}}{R} + 1}{2}, \sigma = \mathrm{diag}(\mathcal{W}_r^2).$$

  - We finally sample k pseudo demonstrations $\boldsymbol{v}$ from k event semantic regions to augment the text generation, that is $\boldsymbol{v} = \{\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}, \ldots \boldsymbol{v}^{(k)}\}$.



Algorithm 1: Gaussian Sampling

Input: The embeddings of schema, document and discrete demonstrations, i.e. $\bar{\mathbf{h}}_{\mathbf{e}}$, $\bar{\mathbf{h}}_{\mathbf{x}}$ and $\mathbf{h_d} = \{\mathbf{h_d}^{(1)}, \mathbf{h_d}^{(2)}, \ldots, \mathbf{h_d}^{(k)}\}$

Output: A set of pseudo demonstrations $\mathbf{v} = \{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(k)}\}$

Normalizing the importance of each element in $\mathbf{b} = \bar{\mathbf{h}}_{\mathbf{x}} - \bar{\mathbf{h}}_{\mathbf{e}}$: $\mathcal{W}_r = \frac{|\mathbf{b}| - \min(|\mathbf{b}|)}{\max(|\mathbf{b}|) - \min(|\mathbf{b}|)}$

while $i \leq (k-1)$ **do**

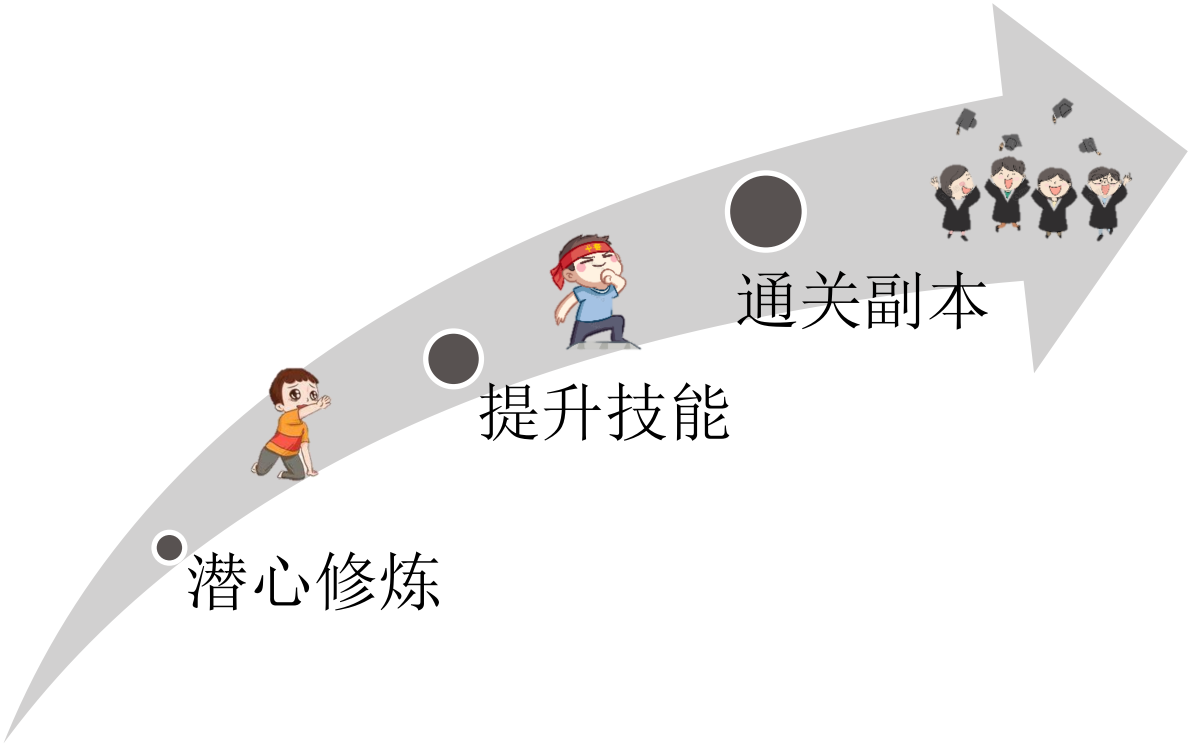
- $i \leftarrow i + 1$
- $r^{(i)} = ||\bar{\mathbf{h}}_{\mathbf{x}} - \mathbf{h}_{\mathbf{d}}^{(\mathrm{i})}||, R = ||\bar{\mathbf{h}}_{\mathbf{x}} - \bar{\mathbf{h}}_{\mathbf{e}}||$
- Use reparametrizetion to calculate the current scale vector: $\omega^{(i)} \sim \mathcal{N}\left(\frac{1 - r^{(i)}/R + 1}{2}, \mathrm{diag}\left(\mathcal{W}_r^2\right)\right)$.
- First sample a noise variable ϵ from $\mathcal{N}(0, 1)$
- Then transform it to $\omega^{(i)} = \mu + \epsilon \cdot \sigma$, where $\mu = 1 - \frac{r^{(i)}/R}{2}, \sigma = \mathrm{diag}(\mathcal{W}_r^2)$.
- Calculate the current sample: $\mathbf{v}^{(i)} = \bar{\mathbf{h}}_{\mathbf{e}} + \omega^{(i)} \odot \mathbf{b}$
- $\mathbf{v} \leftarrow \mathbf{v} \cup \mathbf{v}^{(i)}$

end

Outline

- Background
- Our work
- Experience sharing

Experience× progress of growing up√



潜心修炼→提升技能→通关副本

- 前期论文积累+代码积累
  - 写论文前：精读本领域论文，完全了解本领域
  - 写论文中：看与你论文技术相关的优秀论文，参考立意陈述等
  - 写论文后：扩大阅读范围，转变阅读思路，积累词句，学习实验设计
  - 代码：领域内代表工作的数据处理/指标计算、领域外功能模块设计、训练通用trick

- 论文看多了才知道什么是好论文，知道目标才能勇往直前

- 所有付出都会显式or隐式地回报给你
  - 客观被分配的工作，不一定是没有意义的
  - 主管去做的工作，要有计划、有方向性、有针对性的去完成

潜心修炼→提升技能→通关副本

- 投搞一次=打一次副本，副本难度 UP!，个人技能点 UP!
- 每次写论文都力求进步，要比上次写的更好
  - AAAI2022(×)→COLING2022(√)→ACL2023(√)

    初窥门径 → 略有小成 → 驾轻就熟
- 每次投搞在能力范围内做到最好，把可能导致拒稿的客观原因都kill掉
  - 写作重点为Introduction：把故事讲清，在此基础上把故事写漂亮
  - 保证solid：包括但不限于对比数据集/baseline要全，引用完整，公式规范，语法错误使不得…
- 写论文or选择会议要有自己的判断，不人云亦云

潜心修炼→提升技能→通关副本

人生是旷野，慢慢往前走就好

祝你好运

祝我好运

Thanks!

renyubing@iie.ac.cn
May. 2023