LREC-COLING 2024

Event-enhanced Retrieval in Real-time Search

Yanan Zhang, Xiaoling Bai, Tianhua Zhou
Tencent Search, Platform and Content Group
{yananzhang, devinbai, kivizhou}@tencent.com

Motivation

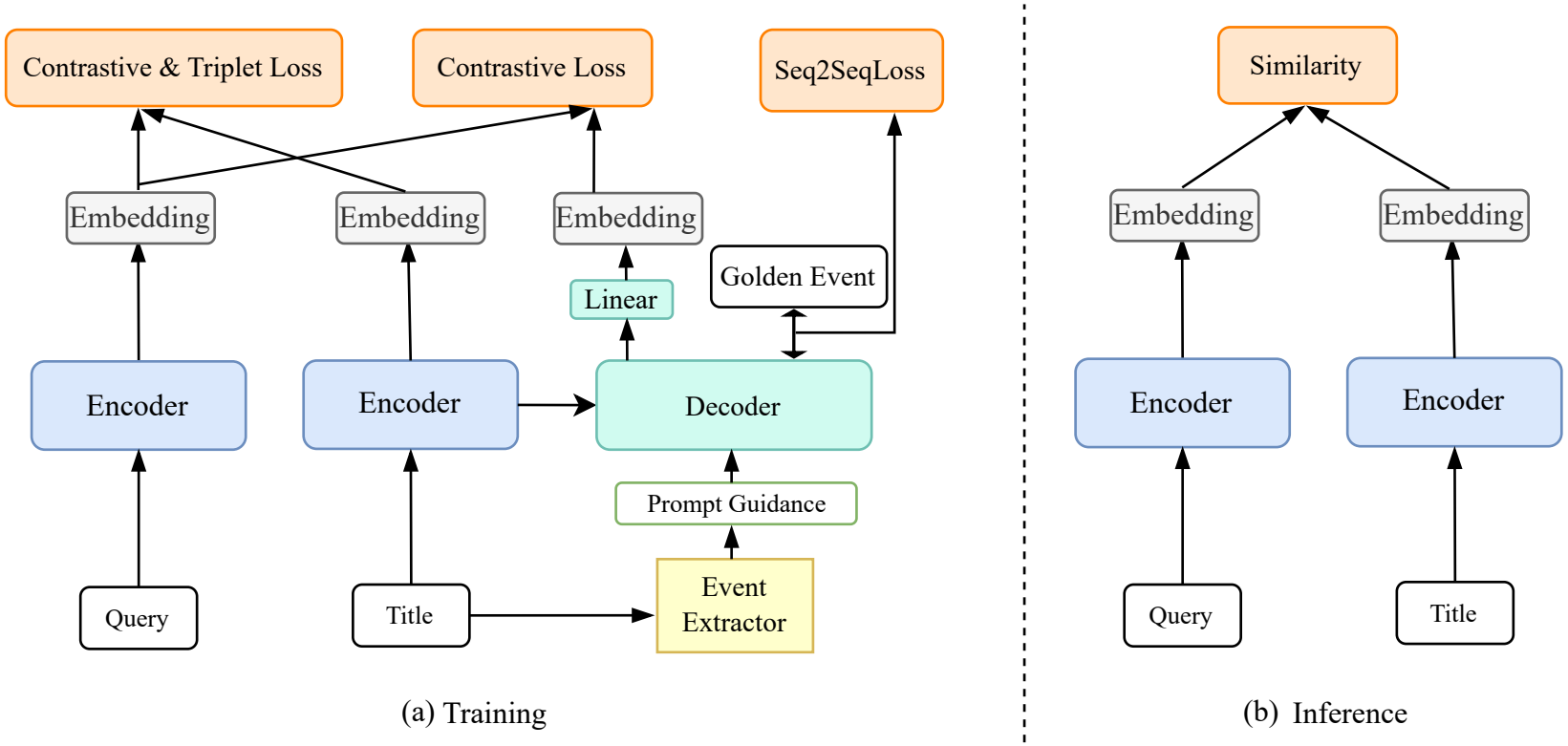
- **Semantic Drift**: The encoded semantics of the model deviate from the user's given context.
- **Real-Time Retrieval**: Users expect to receive up-to-date information about current events.
- **Information Asymmetry**: A significant difference in the amount and type of information contained in user queries and document titles.
- **Event Focus**: The importance of capturing event-related information is paramount.

Event: 华为Mate60 Pro开售 (Huawei Mate60 Pro goes on sale)
Different Queries
1. 华为Mate60pro (Huawei Mate60pro)
2. mate60pro (mate60pro)
3. mate60pro价格 (mate60pro price)
4. 华为Mate60pro上线 (Huawei Mate60pro goes online)
5. mate60pro咋样 (How about mate60pro)
6. 华为Mate60pro对比 (Huawei Mate60pro comparison)
7. 华为mate60pro 最新消息 (Huawei mate60pro latest news)
8. Mate60pro有耳机吗 (Does Mate60pro have headphones?)
Different Document Titles
1. 华为不讲“武德”？6999元开售Mate60，全球第一款卫星通话的手机 (Huawei doesn’t respect “martial ethics”? Mate60, the world’s first satellite phone phone, goes on sale for 6,999 yuan.)
2. 华为Mate 60 Pro悄然发布!这些规格参数很亮眼，快来一睹为快吧 (Huawei Mate 60 Pro was quietly released! These specifications are very eye-catching, come and take a look)
3. 稳了!6999元,华为Mate60 Pro震撼回归!你的下一部梦幻手机已经诞生! (Stable! At 6,999 yuan, Huawei Mate60 Pro makes a shocking return! Your next dream phone has been born!)
4.入手华为mate60pro！*#华为mate60pro+##华为mate60开售##遥遥领先#* (Get Huawei mate60pro! *#huaweimate60pro+##huawei mate60 is on sale##way ahead#)*

Our Contribution

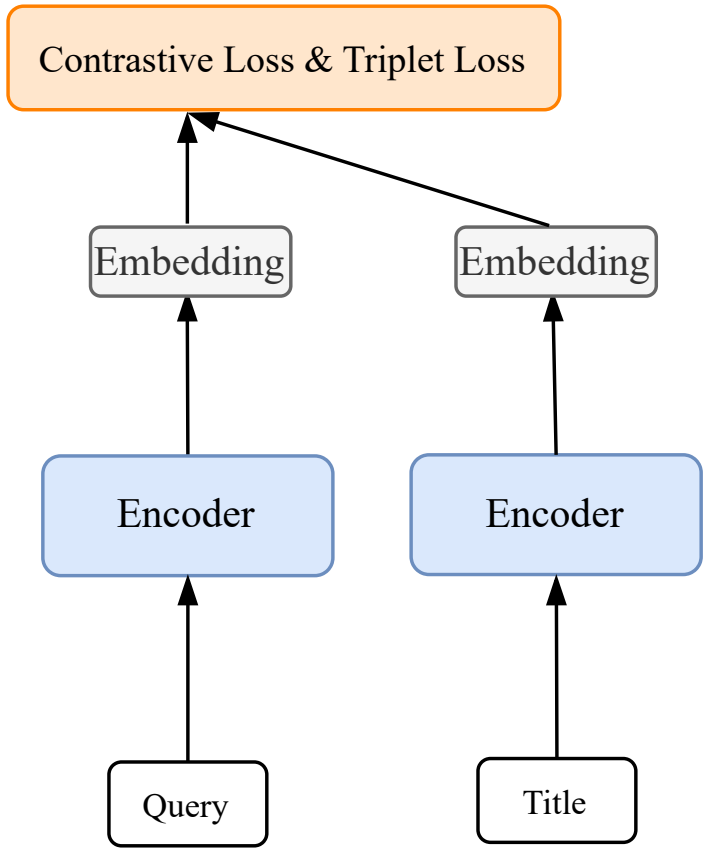
- A new event-enhanced retrieval method to boost real-time search systems, especially by tackling semantic drift and highlighting key event data.
- Enhancing the traditional dual-tower model with a specialized event generation task.
- Extensive experiments with a Tencent QQ Browser Search log-based dataset show EER significantly outperforms current method.

Methodology



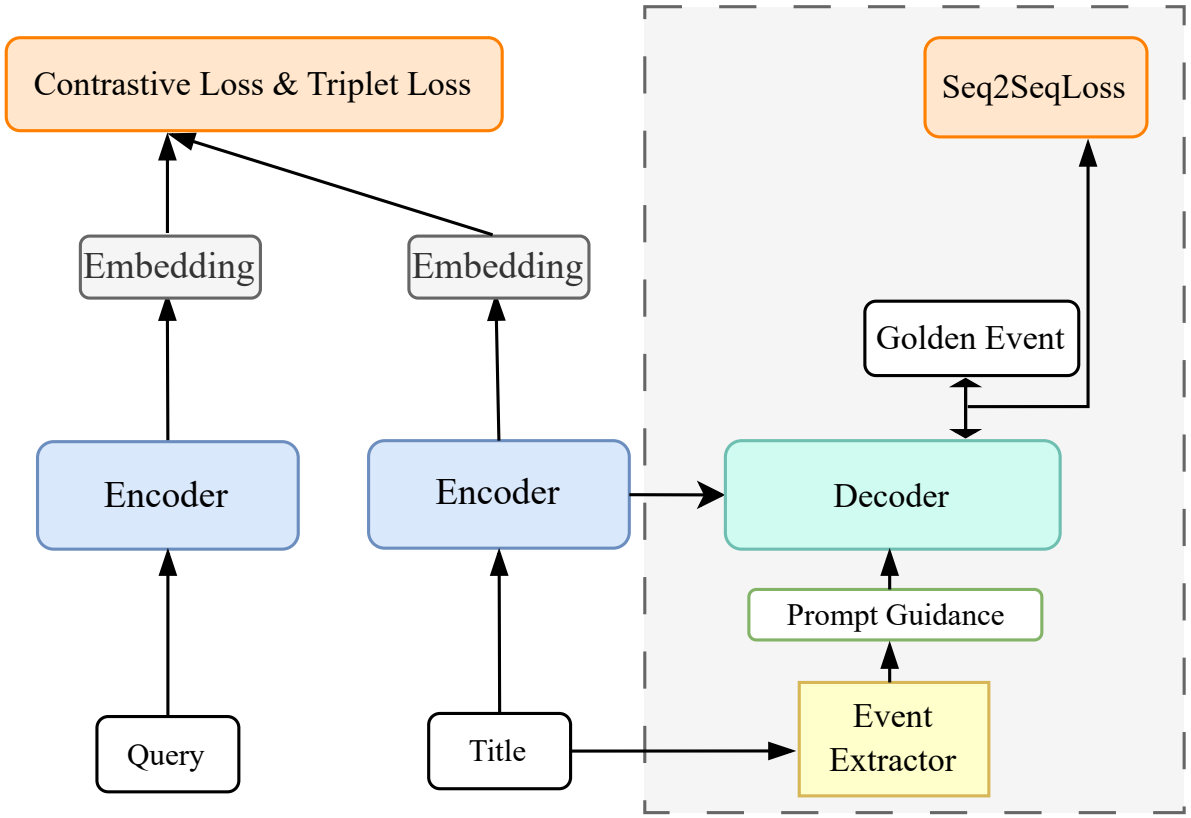
Model Architecture

Methodology



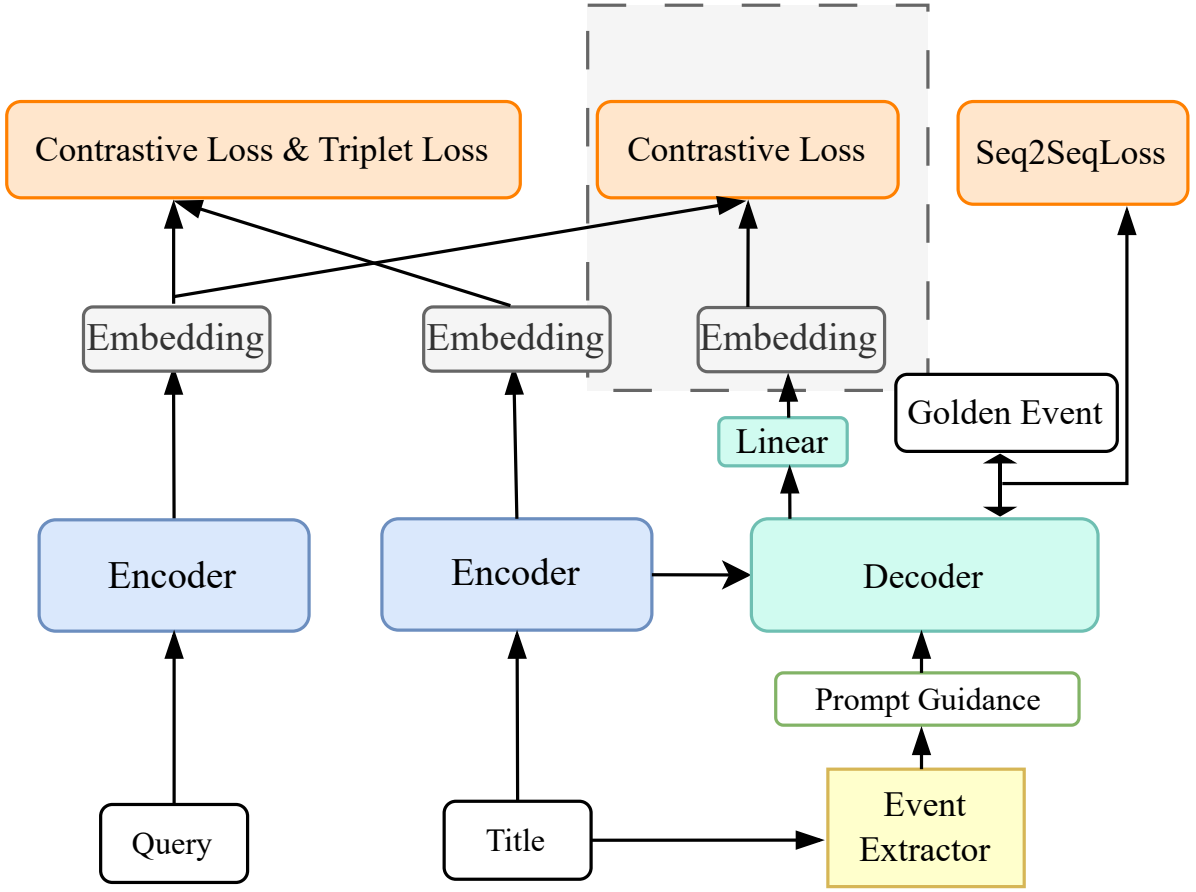
- Hard Negative Sampling
  - Knowledge Augmentation
  - Semantic Mining
- Contrastive Learning
- Pairwise Learning

Methodology



- Generative Decoder
  - Event Extraction
- Prompt Guidance

Methodology

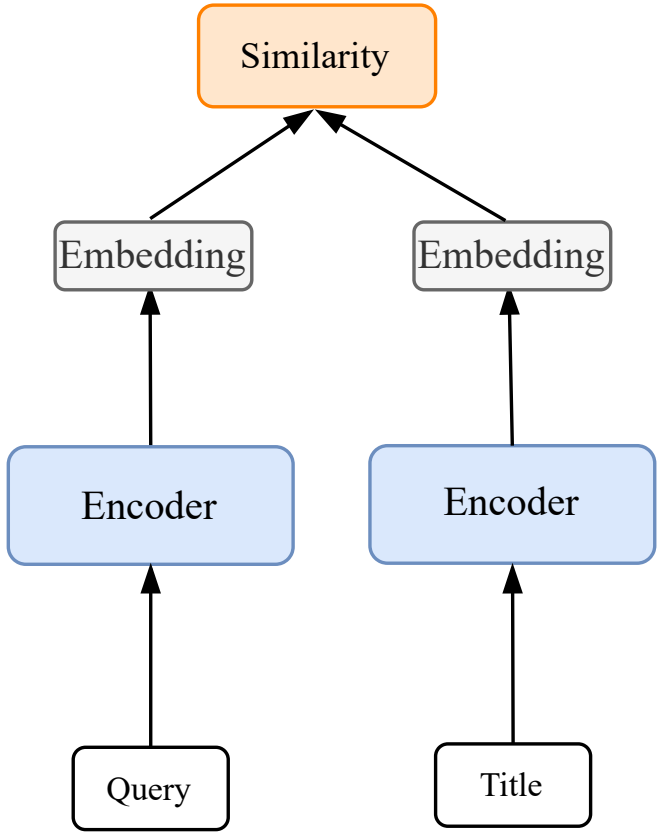


- Relevance Learning Between Query and Event
  - Contrastive Learning

- Total loss

$$\mathcal{L}_{total} = \mathcal{L}_{cl_{qt}} + \mathcal{L}_{pair_{qt}} + \mathcal{L}_{gen} + \mathcal{L}_{cl_{qe}}$$

Methodology



- Inference:

  Reverting to the traditional dual-tower model

Experiments

Dataset: Producing a dataset for real–time retrieva land make it publicly available.

Data Source: Derived from Tencent QQ Browser Search logs and the majority of the data is in Chinese.

Content Types: Including titles from various document types such as text, user–generated content, videos.

News Categories: Spanning across 23 different news categories.

Annotation Process: Using a combination of user behavior metrics and relevance features.

Expert Annotation: Expert annotation to the test data.

Data Leakage Prevention: Using the same sampling method but different timeframes for the training and test datasets.

Dataset	Queries	Titles	Q-T pairs
Training	2,964,077	5,323,681	10,319,501
Testing	1,096	4,733	10,2279

The statistics of the dataset.

Experiments

Results

Models	R@10	MRR@10	AUC
BM25	0.579	0.556	0.773
Sentence-BERT	0.693	0.650	0.827
BGE	0.771	0.694	0.915
EER	**0.829**	**0.757**	**0.931**

Evaluation of EER and baselines.

Models	R@10	MRR@10	AUC
base	0.687	0.597	0.829
base+CL	0.734	0.628	0.850
base+CL+GD	0.769	0.673	0.884
base+CL+GD+GP	0.786	0.679	0.910
base+CL+GD+QER	0.802	0.704	0.915
EER	0.829	0.757	0.931

Evaluation of EER components.

Experiments

Case Study

Case	Sentence-Bert	BGE	EER	Discripsion
Label: 0 Query: 华为mate50 (Query: Huawei mate50) Title: 华为Mate60突然开售，没有任何预告 (Title: Huawei Mate60 suddenly goes on sale without any notice)	label:1	label:1	label:0 Similarity:0.317	Huawei “Mate50” and “ Mate60” are different phone series, so this case is irrelevant.
Label: 1 Query: 日本核废水 (Query: Japanese nuclear wastewater) Title: 定了!日本8月24日将排放福岛核污水,中方坚决反对 (Title: It's decided! Japan will discharge Fukushima nuclear wastewater on August 24, and China firmly opposes it.)	label:0	label:0	label:1 Similarity:0.784	“污水(sewage)” and “废水(wastewater)” are different words but express almost the same meaning and this case is relevant.

Summary

- EER is an embedding-based method to boost real-time search's semantic retrieval
- It uses a generative decoder to better understand implicit event info, improving query matching and reducing semantic drift.
- Extensive tests prove EER's effectiveness, and it's cost-effective as it doesn't change model parameters during inference.

Summary

- EER is an embedding-based method to boost real-time search's semantic retrieval
- It uses a generative decoder to better understand implicit event info, improving query matching and reducing semantic drift.
- Extensive tests prove EER's effectiveness, and it's cost-effective as it doesn't change model parameters during inference.

Thank you