Enhancing Search with Malaysian Reranking

Wan Adzhar Faiq Adzlan[*] Kamarul Adha[†] Husein Zolkepli[‡] Aisyah Razak[§]
Mas Aisyah Ahmad[¶] Halim Shukor[‖]

Abstract

1 Introduction

We present an innovative adaptation to Retrieval Augmented Generation (RAG), aimed at refining the document retrieval process for improved relevance. Traditionally, RAG involves converting input and potential output into embeddings, subsequently sorting them via semantic search. However, to further enhance the retrieval of relevant documents, we introduce a reranker model. This model is trained to assess the similarity between two documents using cross-entropy loss, akin to a classification model. By assigning probability values between 0 and 1 to document pairs, the reranker effectively reorders documents, prioritising relevance.

While several reranker models have been introduced, none have been trained on local Malaysian data. To address this gap, we adopt a strategy of leveraging our own data to train the reranker model. This adaptation allows us to tailor the reranking process to our specific context, enhancing the efficacy of our RAG architecture. By integrating the reranker model alongside the embedding model and semantic search, we present a comprehensive solution for optimising document retrieval in RAG frameworks.

2 Post-Mining and Post-Filtering

2.1 Dataset Preparation

Since a reranker serves to enhance an embedding model and vector search, it necessitates a dataset rich in both highly relevant (strong positive) and highly irrelevant (strong negative) examples. This is pivotal as the reranker's function is to augment document retrieval by discerning the most relevant documents for a given query, thereby surpassing the semantic recognition capabilities of the embedding model.

Our utilization of the pretrained embedding model mesolitica/llama2-embedding-600m-8k facilitates the conversion of our dataset into embeddings. In contrast to prior methodologies [**?**], wherein we employed OpenAI-ada and bge-large for this purpose, the adoption of our current embedding model, which has been trained on data imbued with Malaysian context, stands as a logical choice for post-mining.

Following the conversion of the dataset into embeddings, we proceed with a data mining approach that involves recording distances between document pairs and representing these on a distribution

[*] adzhar.faiq@gmail.com
[†] kamarul.adha360@gmail.com
[‡] husein@mesolitica.com
[§] aisyahrazak171@gmail.com
[¶] masaisyahahmad@gmail.com
[‖] mhalimshukor@gmail.com

Technical Report.

graph. Analysis of this graph enables the determination of upper and lower threshold bounds, guiding the selection of relevant documents.

Documents situated at the extremes of the threshold spectrum represent varying degrees of similarity: those at the higher end exhibit greater likeness, whereas those at the lower end demonstrate dissimilarity. By curating a dataset comprising pairs from both ends, we establish a collection of highly positive (similar) and highly negative (dissimilar) document pairs.

3 Synthetic Title-Context Pair

We leverage a comprehensive dataset comprised of deduplicated Malaysian news articles, summaries, and deduplicated online articles. Our methodology involves the creation of positive pairs, wherein the relationship between titles and their corresponding articles is utilized to generate pairs that exhibit semantic coherence and relevance. To achieve this, we meticulously match titles with their respective articles, ensuring that the generated pairs reflect meaningful associations. In parallel, we employ a contrasting approach to form negative pairs, wherein titles are paired with articles different from their original counterparts. This deliberate juxtaposition enables us to capture nuanced distinctions and refine the model's discernment capabilities.

To create negative pairs, we adopt a strategy where articles with minimal overlap between their titles and content keywords are selected. Specifically, we set a threshold of less than 10% keyword overlap between the title and article for the negative pair selection process, below is the pseudo Python code,

```
import re

def clean(string):
  string = re.sub(’[^A-Za-z ]+’, ’ ’, string.lower())
  string = re.sub(r’[ ]+’, ’ ’, string).strip()
  return string

def overlap(string1, string2):
  l = set([w for w in clean(string1).split() if len(w) > 2])
  r = set([w for w in clean(string2).split() if len(w) > 2])
  return len(l & r) / len(l)

title = ’this is title’
body = ’this is body’
negative = []
if overlap(title, body) < 0.1:
  negative.append(body)
```

All synthetic dataset and implementation published at mesolitica/title-context-pair.

4 Synthetic Noisy Translation

4.1 Finetuning T5 for Noisy Translation

We present an innovative approach to enriching local context social media datasets sourced from platforms including Facebook, Twitter, and b.cari.com.my, c.cari.com.my. The collected data undergoes a crucial preprocessing step where we employ ChatGPT3.5 to translate the content. This process results in what we term as "noisy translation," wherein the translation may not always be perfectly accurate due to the inherent complexities of translating informal and colloquial social media content. Despite the noise introduced during translation, our methodology allows us to effectively bridge language barriers and incorporate diverse linguistic nuances present in the local context data.

An example of noisy translation,

```
{’left’: ’bagusla mawi bagi peluang kat junior2 dia yang berbakat tak kira lelaki
    atau perempuan untuk tonjolkan bakat, bila appear dgn mawi ada la orang
    tertarik nak ambik lagi’,
’en’: "It’s good that Mawi gives opportunities to talented juniors regardless of
    gender to showcase their talents. When they appear with Mawi, there are people
    who are interested in taking them on again.",
’ms’: ’Baguslah Mawi memberi peluang kepada junior-junior yang berbakat tanpa
    mengira jantina untuk menonjolkan bakat mereka. Apabila mereka muncul bersama
    Mawi, terdapat orang yang berminat untuk mengambil mereka lagi.’,
’cleaned’: ’bagusla mawi bagi peluang kat junior2 dia yang berbakat tak kira
    lelaki atau perempuan untuk tonjolkan bakat, bila appear dgn mawi ada la orang
    tertarik nak ambik lagi’}
```

All noisy translation dataset we published at mesolitica/malaysian-noisy-translation.

we introduce a novel approach to enhance the generation of synthetic data with similar semantic meaning but slightly varied sentence structures within the local language context. To achieve this, we leverage the pretrained Malaysian T5 Small model from Malaya [?] to train a reverse noisy translation system. Unlike conventional translation tasks where the focus is on translating from local language to standard language, our methodology involves training the model to translate from standard language to local language. This strategic shift allows us to generate synthetic data that closely mimics the semantics of the original content while introducing subtle variations in sentence structure. By harnessing the power of reverse translation and the capabilities of pre-existing language models, we aim to address the challenge of data scarcity in local language datasets while preserving the linguistic richness and nuances specific to the target language.

Below is the prefix to finetune pretrained Malaysian T5 Small for local Malay language,

```
original = ’bagusla mawi bagi peluang kat junior2 dia yang berbakat tak kira
    lelaki atau perempuan untuk tonjolkan bakat, bila appear dgn mawi ada la
    orang tertarik nak ambik lagi’
ms = ’Baguslah Mawi memberi peluang kepada junior-junior yang berbakat tanpa
    mengira jantina untuk menonjolkan bakat mereka. Apabila mereka muncul bersama
    Mawi, terdapat orang yang berminat untuk mengambil mereka lagi.’
prefix = ’terjemah ke pasar Melayu: ’
input = f’{prefix}{ms}’
output = original
```

4.2 Synthetic Pair

Following fine-tuning, we generate synthetic datasets utilizing Multinomial sampling with specific parameters such as top-p 0.95 and top-k 50, ensuring a balance between diversity and relevance. Moreover, to enhance diversity further, we produce five different outputs for each input sentence along with their corresponding scores. This comprehensive approach not only ensures the authenticity of the synthetic dataset but also promotes diversity.

An example of generated synthetic reversed noisy translation,

```
 {’original’: {’left’: ’Geng 12 hb ni ade tak yang nak balik terengganu ade satu
     ticket dah beli tapi tak jadi pergi nak bagi harga murah j
     https://t.co/15KiWG9vfh’,
 ’en’: "Anyone going back to Terengganu on 12th September? I have a ticket but
     can’t go. Willing to sell at a cheaper price. DM me.",
 ’ms’: ’Ada sesiapa nak balik Terengganu pada 12 September? Saya ada tiket tapi
     tak dapat pergi. Sedia untuk jual pada harga yang lebih murah. DM saya.’,
 ’cleaned’: ’Geng 12 hb ni ade tak yang nak balik terengganu ade satu ticket dah
     beli tapi tak jadi pergi nak bagi harga murah j ’},
’en’: {’score’: [32.828792572021484,
  33.19839859008789,
  32.189056396484375,
  31.707237243652344,
  32.984588623046875],
 ’sequences’: [’sesiapa balik terengganu 12sept takde tiket tak boleh pi nak jual
     harga lagi murah dm me’,
  ’ada yang balik terengganu 12 sept ni ada ticket tapi takleh nak pi bakalan jual
      murah dm sikit’,
  'Anyone going back to Terengganu on 12 September? I got a ticket but cannot go.
      Willing to sell at a cheaper price. DM me.',
  'Anyone going back to Terengganu on 12th September? Got ticket but cant go.
      Willing to sell at a cheaper price. DM me',
  "Anyone going back to Terengganu on 12th September? I have a ticket but can't
      go. Willing to sell at a cheaper price. DM me."]},
'ms': {'score': [32.65056228637695,
  30.710777282714844,
  32.17496109008789,
  33.49098205566406,
  31.255817413330078],
 'sequences': ['Ada sapa nak balik terengganu 12 sept? Ada tiket tapi tak boleh
     pergi. Siap jual lagi murah. DM aku',
  'Ada sesiapa nak balik Terengganu 12 Sep. Aku ada tiket tak dapat. Ready to sell
      harga murah. DM aku',
  'Ada sapa2 nak balik terengganu 12 Sep ni? I ada tiket tapi tak boleh pergi.
      Siap nak jual murah lagi. DM me https://t.co/LhxYzFk7xY',
  'Ada yang nak balik terengganu 12 september? Aku ada tiket tapi tak dapat pergi.
      Siap nak jual harga murah. DM aku',
  'Ada siapa nak balik Terengganu 12 September? Ada tiket tapi tak boleh pergi.
      Ready nak jual murah. DM saya']}}
```

To further promote dataset diversity, we meticulously curate positive and negative pairs. Positive pairs are constructed by selecting translations with an average logprob score of at least 30, combined with noisy translations generated by ChatGPT3.5. Conversely, for negative pairs, we select translations with minimal keyword overlap, specifically less than 5%, to ensure distinctiveness from the positive pairs.

An example of positive and negative pairs,

```
{'negs': ['Rabu / 18 Mei 2022 / 17 Syawal 1443H\n5:51pg - Masuk waktu solat fardhu
     #Subuh bagi Pulau Pinang &amp; kwsn yg sama wakt https://t.co/zfdy4mpC1x',
  '@nilamsaniiiiii @idek_hm ngak g gya ... matik bak isik borang sen agy',
  'Labuan Bajo adalah salah satu destinasi wisata super prioritas dan premium,
      namun kami ingin agar wisatawan nusatar https://t.co/hVzQPVLFLS',
  '@BangRiz91376468 Mau, dong....',
  'gak kasian sama aku ya? oke anak kita nambah https://t.co/WG6a1DyT0g'],
 'pos': ['For your information, last week during Eid al-Fitr, Ms Maharani was known
     to have had a political meeting with the Chairman of https://t.co/E1ITii7Pcl.',
  'Sebagai makluman, minggu lalu semasa raya al-fitri, Puan Maharani diketahui
      telah mengadakan pertemuan politik dengan Pengerusi https://t.co/E1ITii7Pcl',
  'Untuk pengetahuan, pekan lalu saat lebaran, Puan Maharani diketahui telah
      mengadakan pertemuan politik dengan Kepala Maj https://t.co/E1ITii7Pcl',
  'Sebagai informasi, pekan lalu selama lebaran, Puan Maharani diketahui telah
      mengadakan pertemuan politik dengan Pengerusi https://t.co/E1ITii7Pcl.',
  'Untuk makluman, minggu lalu semasa Eid al-Fitr, Puan Maharani diketahui telah
      mengadakan pertemuan politik dengan Pengerusi https://t.co/E1ITii7Pcl',
  'Untuk pengetahuan, minggu lalu selama lebaran, Puan Maharani diketahui telah
      melakukan pertemuan politik dengan Pengerusi https://t.co/E1ITii7Pcl',
  'Sebagai makluman, minggu lalu semasa Eid al-Fitr, Puan Maharani diketahui telah
      mengadakan pertemuan politik dengan Pengerusi https://t.co/E1ITii7Pcl.',
  'For your information, last week during Eid al-Fitr, Puan Maharani was known to
      have had a political meeting with the Chairman of https://t.co/E1ITii7Pcl.'],
 'query': 'Sebagai informasi, pekan lalu selama lebaran, Puan Maharani diketahui
     melalukan pertemuan politik dengan Ketua Umum https://t.co/E1ITii7Pcl'}
```

All synthetic dataset and implementation published at mesolitica/title-context-pair.

5 Finetuning Procedure

We split N train set M test set.

Use Malaysian Mistral 64M, 191M and 474M on binary cross entropy

6 Evaluate

6.1 Post-sorting

6.2 Without Post-sorting

7 Acknowledgement

Special thanks to Malaysia-AI volunteers especially Ammar Azman, M. Amzar, Muhammad Farhan, Syafie Nizam, Alif Aiman, Azwan Zuharimi and Haziq Zikry for contributing dataset to train Malaysian Reranker models.

We would like to express our gratitude to NVIDIA Inception for generously providing us with the opportunity to train our model on the Azure cloud. Their support has played a crucial role in the success of our research, enabling us to leverage advanced technologies and computational resources.

We extend our thanks to the wider research community for their valuable insights and collaborative discussions, which have greatly influenced our work. This paper reflects the collective efforts and contributions from both NVIDIA Inception and the broader research community.

8 Conclusion

Reranker bagi cekang lepas embedding, bagi RAG lagi cekap.