Pretrained Transformers for Text Ranking: BERT and Beyond

Chapter 3 to section 3.2.2

1. Questions

1. **Explicação de conceitos importantes do artigo**
2. A contribuição do artigo
3. **Resultados interessantes/inesperados**
4. **Uma dúvida "básica" que você ou os colegas possam ter**
5. **Um tópico "avançado" para discutirmos**

2. Explicação de conceitos importantes do artigo

1. pré-treinamento gerando eficiência e agilidade para tarefas específicas (**downstream**)
2. BERT - Transformer *encoder-half*
3. GPT - Transformer *decoder half*
4. **self supervision** (sem anotação) no BERT (MLM - *masked language model* bidirecional) e GPT (unidirecional treinado com tokens anteriores prevendo o próximo)
5. limitação tamanho texto BERT
6. tipos diferentes de uso da arquitetura BERT (1+ inputs e outputs usando 1+ vetores de contexto criados)
7. Tokenizadores de sub-palavras - menor vocab
8. vetores de entrada no tranformer que são aprendidos no treinamento e são **somados** não concatenados
    1. **embeddings de tokens** - texto tokenizado
    2. **embeddings de segmentos** - identificar a qual sentença de entrada pertence o token
    3. **embeddings de posição** - identificar posição do token no texto

2.1 Explicação de conceitos importantes do artigo

9. impraticável usar BERT para inferência de milhões de textos para cada *query* - *retrieve-and-rerank architecture*
10. BERT usado como reranquedor nas saídas de um buscador de palavras-chaves com índice invertido
11. **MonoBERT** cross-encoder (combina query e texto em um input)

3. Resultados interessantes/inesperados

- *ablation study* muito esclarecedor
  - alteração do *input template*: ordem deles e remoção de alguns
  - MonoBERT é **data hungry**, precisou de muitos dados pra melhorar bater o BM25 inicialmente
  - combinação polinomial com score do BM25

4. Uma dúvida "básica" que você ou os colegas possam ter

Diferença entre NLP e IR

> O objetivo do NLP é permitir que os computadores processem e entendam a linguagem escrita ou falada, para que possam interagir de forma mais natural e útil com os usuários.

> a IR é a área que se concentra em como os usuários podem encontrar informações relevantes a partir de grandes volumes de dados, como textos ou bancos de dados

- retroalimentação entre as áreas, principalmente nos tempos atuais de deeplearning e tranformers.
- **MonoBERT** menciona que o texto da query é usado "verbatim" do usuário ou da coleção de teste. **DÚVIDA ?**
- Porque o BERT sem o embedding de posição trataria o input como se fosse um **bag of words** ?

5. Um tópico "avançado" para discutirmos

> Hybrid List Aware Transformer Reranking, is a lightweight reranking framework for text retrieval. This framework is premised on combing the retrieval featuring and ranking feature with a list-wise encoder as the reranking model. More details can be found in our

Candidato e artigo interessante. Uma versão dele está sendo usada no 1º lugar do MS MARCO Passage Ranking Leaderboard

https://github.com/AlibabaResearch/HLATR