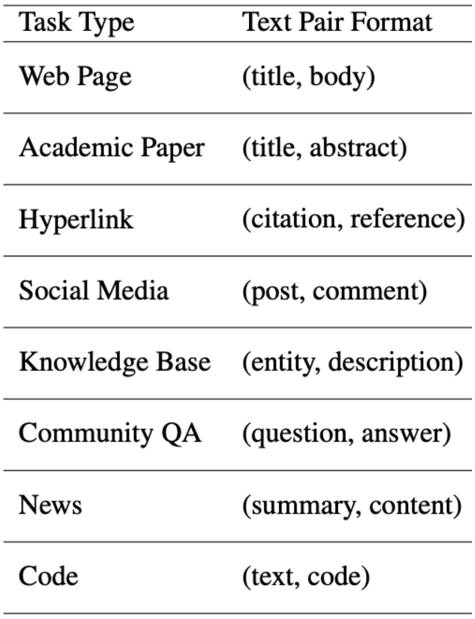
GTE: General Text Embeddings

Link: https://arxiv.org/pdf/2308.03281.pdf

How?

- unsupervised contrastive pre-training and supervised finetuning on BERT based (mean pooling on text representation)
- trained using weak supervised pairs, use open source data without filtering helps in **domain generalization**, in supervised finetuning, prompts are generalised.
- **Impact of scaling the data size is huge** unsupervised data generation, addressing imbalance through multinomial distribution, depends on number of examples, all training examples within a batch are of a same task



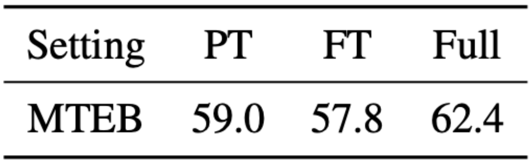
Task Type	Text Pair Format
Web Page	(title, body)
Academic Paper	(title, abstract)
Hyperlink	(citation, reference)
Social Media	(post, comment)
Knowledge Base	(entity, description)
Community QA	(question, answer)
News	(summary, content)
Code	(text, code)

- supervised finetuning is human annotated and less in number,
- improved contrastive learning, **large number of negatives, within batch and documents itself.**

$$
\begin{aligned}
Z = \sum_j e^{s(q_i,d_j)/\tau} + \sum_{j \neq i} e^{s(q_i,q_j)/\tau} \\
+ \sum_j e^{s(q_j,d_i)/\tau} + \sum_{j \neq i} e^{s(d_j,d_i)/\tau}
\end{aligned}
$$

first two terms are for query document contrast and last two terms are inverse. where s(q,d) is cosine similarity

- GTE small is minilm uncased, rest is bert based, in pretraining only in batch negatives is used with large batch size, in finetuning because of negatives a small batch size works as it helps in gradient estimation ⇒ (q,1+,n-)
- Evaluation, text classification as similarity problem, text embedding, label for classification embedding, get similarity, two verbalizers, **prompt of ‘this is negative’ or just word as negative,**
- In code search, model is better than model that is finetuned for a specific structure (code, lang), with larger data, better representations are captured,



Setting	PT	FT	Full
MTEB	59.0	57.8	62.4

pt is just unsupervised, ft is supervised, full is multi-stage, unsupervised pt with supervised ft

Why and related word?

- SimCSE is bad at retrieval since it is symmetric.
- pretrained language models might not give high quality embeddings due to the presence of **unstable embedding** spaces resulting from the MLM objective
- some objectives are through constructing positive pairs by random passage cropping.
- construction of unified text representation models through large-scale contrastive learning and prompt-based learning

Advantage

- performs well in code retrieval too