Information Retrieval System Analysis Report

Team Information

Members:

1. Victoria (Tzu-Ying) Cheng
    - Implemented Part 1: Document Ranking with different distance metrics
    - Conducted experiments on preprocessing impact
    - Analyzed ranking behavior and wrote report for Part 1
2. Rui Tao
    - Implemented Part 2: Fine-tuning experiments with different strategies
    - Conducted experiments with various loss functions and training parameters
    - Analyzed model performance and wrote report for Part 2

Part 1: Ranking Documents Report (10 Points)

Comparison of Encoding Methods

1. GloVe embeddings vs. Sentence Transformer embeddings

- Sentence Transformer MAP score: 0.4774
- GloVe embeddings MAP score: 0.0884
- Sentence Transformer performed substantially better, with a MAP score more than 5 times higher

2. Which method ranked documents better?

- Sentence Transformer clearly outperformed GloVe embeddings
- The contextual embeddings from Sentence Transformer showed superior ability in capturing semantic relationships
- The pre-trained model demonstrated strong out-of-the-box performance for this domain

3. Did the top-ranked documents make sense?

Using the example query "Breast Cancer Cells Feed on Cholesterol":

- Top documents (MED-2439, MED-2434) showed high relevance with similarity scores 0.69-0.67
- The rankings appeared logically consistent with the query topic
- All top 10 results maintained meaningful similarity scores above 0.50

4. How does cosine similarity behave with different embeddings?

- Sentence Transformer: Produced well-distributed similarity scores (0.69 to 0.50)
- GloVe: Generated lower and more compressed similarity ranges
- Sentence Transformer showed better discrimination between relevant and irrelevant documents

Observations on Cosine Similarity & Ranking

1. Did the ranking appear meaningful?

- Yes, the similarity scores decreased gradually
- Clear differentiation between more and less relevant documents
- The distribution of scores suggested good discrimination ability
- Additional experiments with different queries confirmed consistent ranking behavior

2. Were there cases where documents that should be highly ranked were not?

- Some domain-specific relationships might have been missed
- Technical medical terms could affect ranking accuracy
- Context-dependent relevance might not always be captured
- Experiments with different distance metrics showed similar patterns of missed relevant documents

3. What are possible explanations for incorrect rankings?

- Lack of domain-specific training in the base model
- Complex medical terminology relationships
- Absence of document structure consideration in the ranking
- Different distance metrics (Euclidean, Manhattan) showed similar limitations in capturing certain relationships

Observations on Distance Metrics & Ranking

1. Comparison of Different Distance Metrics

Experimental results with the query "Do Cholesterol Statin Drugs Cause Breast Cancer?":

- Cosine Similarity:
  - Found 5/10 relevant documents in top 10

    - Score range: 0.7492 to 0.5948
    - Top relevant documents: MED-2429, MED-10, MED-2431, MED-14, MED-2428
    - Better at capturing semantic relationships
- Euclidean Distance:
    - Also found 5/10 relevant documents in top 10
    - Distance range: 0.7083 to 0.9002
    - Same top relevant documents as cosine similarity
    - Similar ranking pattern but different score distribution
- Manhattan Distance:
    - Found 6/10 relevant documents in top 10
    - Distance range: 11.1150 to 13.9932
    - Additional relevant document: MED-4559
    - More sensitive to term-specific differences

2. Were there differences in ranking behavior?

- All metrics identified the same top 5 relevant documents
- Manhattan distance found one additional relevant document (MED-4559)
- The ordering of less relevant documents varied between metrics
- Score distributions showed different characteristics but maintained similar effectiveness

3. Impact of Preprocessing

Experiments with basic preprocessing (lowercase, punctuation removal):

Original Query Results:

- Query: "Do Cholesterol Statin Drugs Cause Breast Cancer?"
- Found 6/10 relevant documents in top 10
- Score range: 0.7492 to 0.5948
- Relevant documents not in top 10: 18
- Examples of missed relevant: MED-2427, MED-2430, MED-2432

Preprocessed Query Results:

- Query: "do cholesterol statin drugs cause breast cancer"
- Also found 6/10 relevant documents
- Score range: 0.7977 to 0.6238
- Similar pattern of relevant document distribution
- Same missed relevant documents

Possible Improvements

1. What can be done to improve document ranking?

- Implement domain-specific pre-processing
- Use medical domain-adapted models
- Combine multiple embedding methods
- Experiments showed potential benefits of combining different distance metrics

2. Would a different distance metric help?

- Euclidean distance showed comparable performance (5/10 relevant in top 10)
- Manhattan distance slightly better (6/10 relevant in top 10)
- Hybrid similarity measures could improve ranking quality
- Each metric showed unique strengths in document discrimination

3. Would preprocessing improve ranking?

- Basic preprocessing (lowercase, stopwords) showed minimal impact
- Medical term normalization might help
- Document structure awareness could be beneficial
- Experiments demonstrated Sentence Transformer's robustness to text variations

Part 2: Fine-Tuning Report (15 Points)

Comparison of Different Training Strategies

1. [anchor, positive] vs [anchor, positive, negative]

Results:

- Original model MAP: 0.4774
- [anchor, positive, negative] MAP: 0.4748
- [anchor, positive] MAP: 0.4632

2. Which approach seemed to improve ranking?

- The triplet approach ([anchor, positive, negative]) performed better
- Both approaches showed slight performance degradation from the original model
- Negative samples provided important contrastive learning signals

3. How did the model behave differently?

- Triplet approach maintained better discrimination between relevant and irrelevant documents
- Pair approach showed less stable ranking behavior
- Negative samples helped maintain ranking quality

Impact on MAP Score

1. Did fine-tuning improve or hurt the MAP score?

Training duration impact:

- 5 epochs: 0.4774 -> 0.4767
- 10 epochs: 0.4774 -> 0.4769
- 20 epochs: 0.4774 -> 0.4737

Negative samples impact:

- 1 negative: 0.4774 -> 0.4743
- 3 negatives: 0.4774 -> 0.4690
- 5 negatives: 0.4774 -> 0.4690

2. Why might MAP have decreased?

- Longer training (20 epochs) led to slight performance degradation
- More negative samples (3-5) didn't improve performance
- The model was already well-optimized for the task
- Pre-trained model's strong baseline performance was difficult to improve upon

3. Is fine-tuning always necessary?

- Not always necessary with strong pre-trained models
- Benefits depend on domain similarity to pre-training data
- Cost-benefit analysis needed for specific applications

Observations on Training Loss & Learning Rate

1. Did the loss converge?

- Training completed in approximately 1 hour for each run
- MultipleNegativesRankingLoss showed stable convergence
- ContrastiveLoss showed steady but suboptimal convergence
- TripletLoss exhibited unstable training behavior

2. Was the learning rate appropriate?

- Default learning rate with warmup strategy proved effective
- Initial rapid loss decrease suggests appropriate choice
- Steady convergence without oscillations indicates good learning rate

3. How did freezing/unfreezing layers impact training?

- Unfrozen model (all layers trainable):
  - Better performance (MAP: 0.4965)
  - Longer training time (~1:07:05)
  - More adaptability to domain
- Frozen layers except last:
  - Slightly lower performance (MAP: 0.4748)
  - Faster training
  - Limited model adaptability

Future Improvements

1. Would training with more negatives help?

Based on our experiments:

- Increasing negative samples did not improve performance:
  - 1 negative: MAP decreased to 0.4743
  - 3 negatives: MAP decreased to 0.4690
  - 5 negatives: MAP remained at 0.4690
- Future work could explore more sophisticated negative mining strategies

2. Would changing the loss function improve performance?

Our experiments with different loss functions showed:

- MultipleNegativesRankingLoss: Best performance (0.4748)
- ContrastiveLoss: Significant degradation (0.2926)
- TripletLoss: Largest degradation (0.1423)
- Future work could explore custom loss functions for medical domain

3. Could increasing epochs improve the model?

Our experiments with different training durations showed:

- 5 epochs: MAP slightly decreased to 0.4767
- 10 epochs: MAP slightly decreased to 0.4769
- 20 epochs: MAP decreased further to 0.4737
- Future work could explore different learning rate schedules