# Consistency and Variation in Kernel Neural Ranking Model

Mary Arpita Pyreddy*†, Varshini Ramaseshan*†, Narendra Nath Joshi*†, Zhuyun Dai†, Chenyan Xiong†, Jamie Callan† and Zhiyuan Liu‡

Language Technologies Institute, Carnegie Mellon University†, Tsinghua University‡

## Research Question

- How much consistency and variance is there in K-NRM results?
- Non-convexity and stochastic training raises questions about consistency of neural models as compared to heuristic and learning-to-rank models.

## K-NRM

- K-NRM learns the word embeddings and ranking model from relevance signals.
- Its effectiveness is due to
  - **word embeddings** tailored for search tasks
  - **kernels/soft-bins** that group word pairs based on their similarity.
  - **learning-to-rank model** which combines the kernels based on their importance.

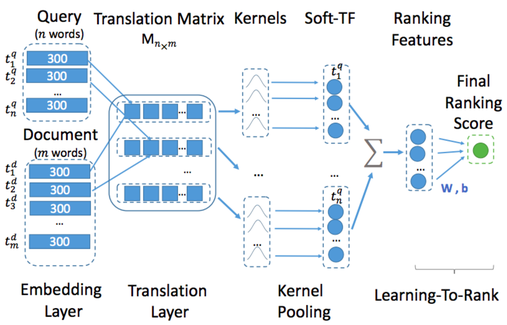



Figure: The architecture of K-NRM

## Sources of Variance in K-NRM

- Random initialization of word embeddings not available in the pre-trained vocabulary.
- Random initialization of weights in the learning to rank layer.

## Experimental Setup

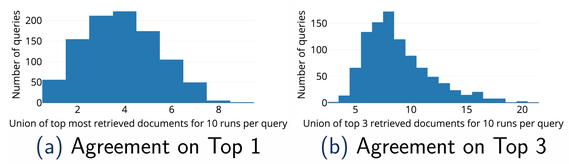
- The consistency of K-NRM was studied by running 50 stochastically trained models with random initialization.
- **Data:** Click log data from Sogou.com, a Chinese web search engine.

## Variance

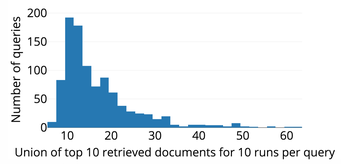
| | Testing-DIFF | | |
|---|---|---|---|
| Statistic | NDCG@1 | NDCG@3 | NDCG@10 |
| Minimum | 0.2983 | 0.3234 | 0.4257 |
| Mean | 0.3242 | 0.3365 | 0.4378 |
| Maximum | 0.3484 | 0.3532 | 0.4496 |
| Std Dev | 0.0108 | 0.0076 | 0.0052 |

Table: Statistics from 50 K-NRM trials trained with random parameter initialization.

- **Variance:** small
- **Min/Max:** Large due to outliers




(a) Agreement on Top 1 (b) Agreement on Top 3




(c) Agreement on Top 10

Figure: Query level ranking agreement. The X-axes are the number of distinct documents that appeared in the top K ranking results of 10 K-NRM trials.

## Takeaway

- K-NRM accuracy is quite stable (has low standard deviation) in spite of its random components.
- Statistics for query level agreements at top 1, 3, and 10 documents:
  - Top 1: 5% of the queries select 1 document; 35% select 2-3 different documents.
  - Top 3: 66% of the queries select 3-9 documents.
  - Top 10: The graph shifts to the left; higher agreement.

## Latent Matching Patterns

**Two sources of variance:**

- **Learning-To-Rank weights**

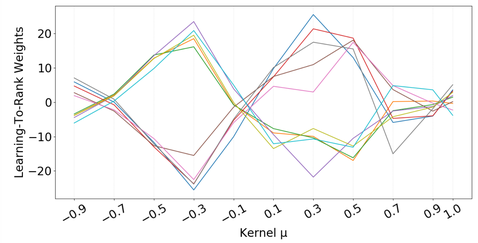
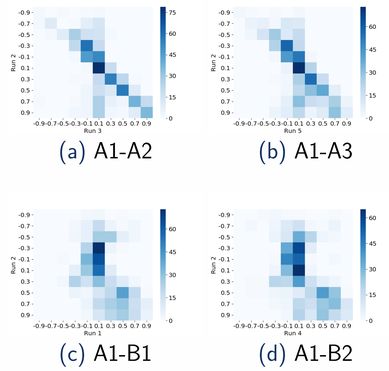



Figure: Learning to rank weights from 10 K-NRM trials.

  - Two distinct patterns A & B are observed in LTR weights from 10 K-NRM trials. Different learning-to-rank weights indicate different ways of allocating word pairs to kernels.
- **Word Embeddings**




(a) A1-A2 (b) A1-A3

(c) A1-B1 (d) A1-B2

Figure: Word pair movements between runs from two patterns, A and B.

- Diagonal relationship is observed in the word pair movements for A1 vs. A2 and A1 vs. A3. Runs from the same pattern have similar learning-to-rank weights and word embeddings, but differ largely in their word pair alignment.

## Takeaway

- Multiple trials of K-NRM converge to two latent patterns that perform similarly.
- Runs within the same pattern converge to similar ranking weights and word embeddings.

## K-NRM Ensembles

**Research Question:** Can knowledge of latent matching patterns enable more accurate ensembles?

- **Ensemble method:** Unweighted average of the document scores.

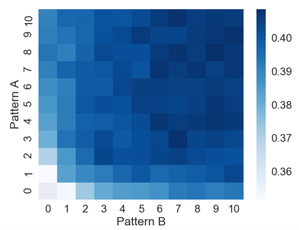



Figure: The MRR of ensemble models that combine different numbers of base models from Patterns A and B.

| | Testing-DIFF | | |
|---|---|---|---|
| Model | NDCG@1 | NDCG@3 | NDCG@10 |
| K-NRM Mean | 0.324 | 0.337 | 0.438 |
| Ensemble-A | 0.370 (14%) | 0.369 (10%) | 0.457 (4%) |
| Ensemble-B | 0.383 (18%) | 0.375 (11%) | 0.463 (6%) |
| Ensemble-A&B | 0.393 (21%) | 0.384 (14%) | 0.468 (7%) |

- Ensemble models are generated from 10 random K-NRM runs.
  - **Ensemble A:** Pattern-A models
  - **Ensemble B:** Pattern-B models
  - **Ensemble A&B:** Pattern-A & Pattern-B models

## Takeaway

- Ensembles that cover an even mix of both patterns are most effective.
- Knowledge of convergence patterns produces more effective ensembles.

## Conclusion

- **Stability:** Accuracy is quite stable, however different trials have moderate agreement about which document to rank first.
- **Latent Matching Patterns:** Multiple trials of K-NRM converge to two latent patterns that are about equally effective.
- **Ensemble:** The distinct but equally effective matching patterns makes K-NRM a good fit for ensemble models.