A better way to format your document for CEUR-WS

Ayush Hirdani, Bhavan Bhatt, Samarth Motka and Harshneel Soni

1. Problem

Ad-hoc information retrieval (IR) involves retrieving a ranked list of documents relevant to a given query from a large corpus. One persistent challenge in this domain is the vocabulary mismatch problem, where semantically related terms in queries and documents differ lexically, leading to poor retrieval performance. Standard retrieval models, such as the Language Model (LM) and its variants, assume term independence, which exacerbates this issue by failing to capture semantic relationships between terms. To address this limitation, we have implemendted a Generalized Language Model (GLM) [1] that integrates word embeddings to model term dependencies explicitly. Unlike traditional approaches, GLM leverages vector-based semantic representations to account for the contextual similarity of terms, enabling improved handling of lexical and semantic variability in queries and documents.

2. Dataset

The experimental evaluation was conducted using the Robust04 dataset[2], a widely-used benchmark for IR tasks. This dataset contains:

- Queries: 250 distinct queries covering diverse topics.
- Documents: Approximately 528,000 documents from the TREC corpus.
- Relevance Judgments: An average of 1245.65 documents per query, with 69.65 documents per query deemed relevant based on relevance judgments.

This dataset's size and diversity make it well-suited for evaluating the robustness and effectiveness of advanced IR models like GLM.

3. Evaluation Metrics

To measure the performance of our proposed GLM, we employed the following metrics:

- Mean Average Precision (MAP): A widely-used metric for evaluating the ranking quality of retrieved documents, with higher MAP values indicating better precision at top ranks.
- Geometric Mean Average Precision (GMAP)[3]: This metric assesses the robustness of a retrieval model across queries by penalizing poor performance on difficult queries.
- Recall: Measures the proportion of relevant documents retrieved by the model, which is critical for understanding its comprehensiveness in addressing vocabulary mismatch.

Introduction to Information Retrieval (IT550), Autumn 2024, DAIICT, India
202101139@daiict.ac.in (A. Hirdani); 202311021@daiict.ac.in (B. Bhatt); 202311023@daiict.ac.in (S. Motka); 202311024@daiict.ac.in (H. Soni)
202101139 (A. Hirdani); 201311021 (B. Bhatt); 202311023 (S. Motka); 202311024 (H. Soni)


4. Methodolody and Results

The GLM introduces two key extensions to the traditional LM framework:

- Term Transformation Events: Instead of directly sampling query terms from documents or the collection, GLM models the generation of a query term as a sequence of events involving the transformation of intermediate terms. These transformations are defined as:
  - Direct Term Sampling: Traditional LM sampling without transformation.
  - Transformation via Document Sampling: Sampling a term from the document and transforming it to the query term using a noisy channel model.
  - Transformation via Collection Sampling: Sampling a term from the entire corpus and transforming it to the query term through the same noisy channel.
- Embedding-Driven Transformations: Word embeddings (e.g., Word2Vec and RoBERTa) are employed to compute semantic similarities between terms. Cosine similarity between vector representations determines transformation probabilities, enabling GLM to account for the contextual fit of terms within documents and mitigate vocabulary mismatches.

Method	MAP	GMAP	Recall
LM	0.2010	0.0943	0.6231
Pre-trained Word2Vec	0.2159	0.1176	0.6305
Word2Vec	0.2531	0.1452	0.6419
RoBERTa	0.3182	0.1927	0.6964

The GLM's performance was compared against the standard LM and LDA-smoothed LM across multiple query sets. Key findings include:

- MAP: GLM consistently outperformed both baselines, with improvements of 5% over LM. The embedding-driven term transformations significantly improved the ranking of semantically relevant documents.
- GMAP: GLM achieved higher GMAP values, demonstrating robustness across queries, including challenging ones with sparse relevance judgments.
- Recall: While recall gains were modest compared to MAP improvements, GLM consistently retrieved more relevant documents than LM.

5. Key challenges and Learnings

Implementing and optimizing the GLM involved several challenges. Balancing the contextual fit of terms within documents and expanding the vocabulary space to include semantically related terms required careful tuning of transformation probabilities. Parameters controlling direct sampling (λ), document-based transformations (α), and collection-based transformations (β) were empirically optimized to maximize retrieval effectiveness. Incorporating word embeddings and computing term transformations necessitated efficient indexing and pre-computation strategies, as storing and accessing nearest neighbor terms for each vocabulary word posed significant scalability challenges. Furthermore, the performance of GLM was sensitive to the weights assigned to the term generation events, with optimal settings varying across datasets and query sets, highlighting the need for adaptive parameter tuning. While embeddings provided a mechanism for modeling term dependencies, interpreting the impact of specific transformations on retrieval outcomes was non-trivial.

References

[1] D. Ganguly, D. Roy, M. Mitra, G. J. Jones, Word embedding based generalized language model for information retrieval, Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (2015). URL: https://api.semanticscholar.org/CorpusID:18739283.

[2] E. M. Voorhees, The trec robust retrieval track, SIGIR Forum 39 (2005) 11–20. URL: https://doi.org/10.1145/1067268.1067272. doi:10.1145/1067268.1067272.

[3] E. Voorhees, Overview of the trec 2004 robust retrieval track, 2005.