ESAQueryRank: Ranking Query Interpretations for Document Retrieval Using Explicit Semantic Analysis

Avijeet Shil
Dept of Computer Science
and Engineering
University of North Texas
3940 N Elm St, Denton, TX 76207
avijeet.shil@my.unt.edu

Wei Jin
Dept of Computer Science
and Engineering
University of North Texas
3940 N Elm St, Denton, TX 76207
weijin@unt.edu

Abstract

Representing query translation into relevant entities is a critical component of an information retrieval system. This paper proposes an unsupervised framework, **ESAQueryRank**, designed to process natural language queries by mapping n-gram phrases to Wikipedia titles and ranking potential entity and phrase combinations using Explicit Semantic Analysis. Unlike previous approaches, this framework does not rely on query expansion, syntactic parsing, or manual annotation. Instead, it leverages Wikipedia metadata—such as titles, redirects, disambiguation pages to disambiguate entities and identify the most relevant ones based on cosine similarity in the ESA space. **ESAQueryRank** is evaluated using a random set of TREC questions and compared against a keyword-based approach and a context-based question translation model (CBQT). In all comparisons of full category types, **ESAQueryRank** consistently shows better results against both methods. Notably, the framework excels with more complex queries, achieving improvements in Mean Reciprocal Rank (MRR) of up to 480% for intricate queries like those beginning with "Why," even without explicitly incorporating the question type. These results demonstrate that **ESAQueryRank** is an effective, transparent, and domain-independent framework for building natural language interfaces.

1 Introduction

Following the COVID-19 pandemic, global reliance on digital platforms for information, education, and remote services has grown significantly. Users now submit complex natural language queries covering diverse topics while expecting accurate and semantically relevant results from shorter queries. This shift has exposed the limitations of traditional keyword-based search models. Early search engines used the bag-of-words model, treating documents and queries as unordered keyword sets. While effective for exact matches, this approach lacked semantic understanding, context, and word order—key elements in natural language interpretation. To address these challenges, various research has explored structured external knowledge sources like Wikipedia. For example, the CBQT framework (Wang et al., 2024) extracts and connects entities via Wikipedia's link structure. However, it relies heavily on graph traversal and does not explicitly model semantic coherence.

In this work, we argue that mapping queries to Wikipedia entities is a powerful tool for building more efficient and context-aware information retrieval systems. For instance, a customer support system where users ask queries like "What is the refund policy?" If the system can understand that "refund policy" is linked to a specific article, it can return more relevant and contextually precise answers. Similarly, in medical applications, where user queries can address diseases, or medications, the ability to map queries to entities like "Paracetamol" can enable more accurate responses, improving user satisfaction and operational efficiency.

Inspired from these, we propose ESAQueryRank, a fully unsupervised, concept-based approach using Explicit Semantic Analysis (ESA). Unlike prior work that merely relied on different effective query expansion techniques(Gupta and Dixit, 2023), ESAQueryRank generates and ranks query-term phrases, selecting the most semantically coherent interpretation within a high-dimensional Wikipedia concept space. The framework maps n-grams with wikipedia articles, leveraging cosine similarity within the ESA space to rank potential entity-phrase combinations. The step filters n-grams that does not represent Wikipedia entity.

After mapping the query n-grams to Wikipedia entities, we build all non-overlapping query-term combinations from relevant phrases. Each query-

Proceedings of Recent Advances in Natural Language Processing,pages 1148–1152
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-132

term combination is then converted in semantic vector using Explicit Semantic Analysis and ranked by cosine similarity to relevant Wikipedia semantic interpretation vectors associated with matched Wikipedia entities. The selection of best query-term combination is based on the highest score. This is worth mentioning that each query-term combination include n-grams that represents one of Wikipedia entities. Based on the highest ranked query-term combination, we feed the combination to an independent external search engine.

To evaluate our approach, we use TREC QA dataset and demonstrate that it consistently shows better results over standard keyword-based baselines and heuristic graph approaches in terms of wikipedia entity resolution accuracy and Mean Reciprocal Rank (MRR).

2 Related Work

Our work bridges the areas of semantic query interpretation, entity disambiguation and linking, concept-based representation via ESA.

Entity Disambiguation and Linking. Entity disambiguation has seen a range of approaches, from supervised learning (Pons et al., 2024; Liu and Nag, 2025) to large-scale, weakly supervised linking (Logeswaran, 2019). Recent efforts incorporate language models like BERT for contextualized entity linking (Wu, 2020; Yan et al., 2025). While effective, these models often require large training datasets and lack interpretability. In contrast, ESA offers a transparent, domain-independent alternative that requires no supervised training.

Concept-Based Representation via ESA. ESA (Gabrilovich and Markovitch, 2007) introduced a sparse, interpretable semantic space using Wikipedia concepts. Although neural embeddings like graph embedding (Ye et al., 2022) or BERT (Devlin et al., 2019; Tashu et al., 2025) have become dominant, ESA remains attractive for systems where concept-level reasoning and interpretability are required. For example, Potthast et al. (Potthast et al., 2008) applied ESA to demonstrate its robustness across different domains. In our work, it enables similarity computation between user queries and Wikipedia entities, providing an interpretable and domain-independent framework for semantic matching.

Semantic Query Interpretation. Work in query understanding has shifted from lexical to semantic modeling. Approaches such as query rewriting using large language models (Kazi et al., 2025; Liu and Mozafari, 2024) and context-aware reformulation (Li et al., 2021) often rely on dense neural retrieval. Our method avoids query rewriting altogether and provides an interpretable alternative to deep learning-based approaches.

3 ESAQueryRank

Figure 1 shows the end-to-end approach involved in *ESAQueryRank* for interpreting natural language questions through semantic matching.

3.1 Wikipedia Entity Matching and Disambiguation Handling

The process begins with processing user query to extract all possible n-grams up to a predefined maximum length (typically, four). Each n-gram candidate is then matched against wikipedia title/entity.

The system incorporates redirect pages and disambiguation page titles (DPTs). If any page refers to any potential redirected entity, we considered the redirected entity for later computation, as the redirected entity contains the most relevant information about the potential entity. Redirects are recursively resolved to their canonical titles. For a disambiguation page, the content of the entity is a list of entities that may match the entity in the common context. For example, in “Mercury (disambiguation),”[1] the entity refers to multiple entities in different contexts, such as a planet, or a chemical component. To resolve, for phrases matching a disambiguation page, all linked entities are extracted, and ambiguity is resolved using TF-IDF-based cosine similarity. We compute the cosine similarity between the semantic vector of the user query, for our case, the question, and the vector representation of each candidate article. The most semantically similar Wikipedia entity is retained.

In some cases, a page may serve both as a redirect and linked in a disambiguation page, based on the context. For example, “United States” is the canonical entity for “US” and “USA,” and is also linked from “America (disambiguation).” These are resolved by recursively identifying the final canonical entity while retaining all potentially linked wikipedia entities for disambiguation. The system ensures and handles hybrid cases effectively while maintaining both correctness and coverage.

[1] `https://en.wikipedia.org/w/index.php?title=Mercury_(disambiguation)`

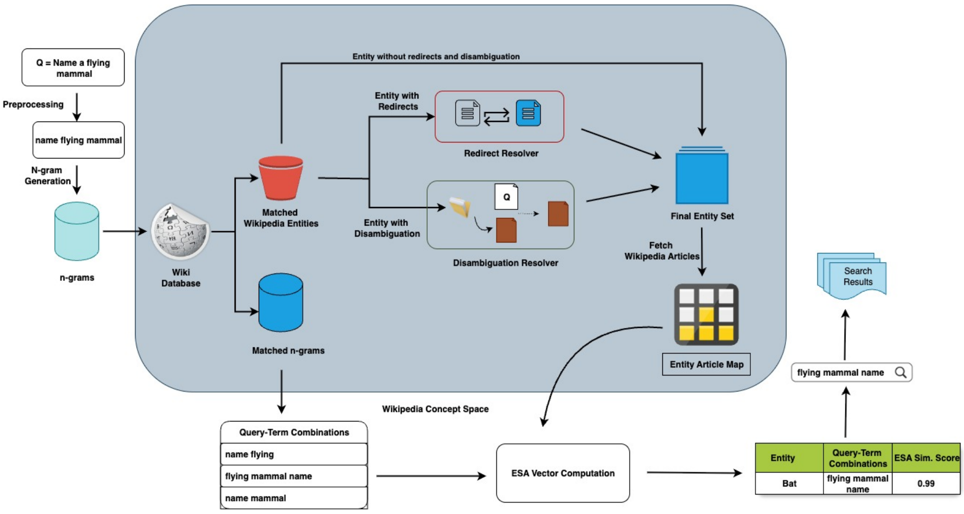

Figure 1: ESAQueryRank Flow for Processing Natural Language Queries.

3.2 Query Interpretation Generation

A query interpretation, in turn, is defined as a set of non-overlapping query-term combinations that together capture the semantic meaning of the original query. For instance, the query "Name a film that has won the Golden Bear in the Berlin Film Festival" may yield combinations such as {*berlin film festival, Golden Bear*} or {*berlin film festival, won*}. Each query-term combination is treated as a distinct candidate for query interpretation. To determine which combination most accurately reflects the user's intent, we use ESA-based representation.

3.3 ESA-Based Semantic Representation

Explicit Semantic Analysis (ESA)(Gabrilovich and Markovitch, 2007) is a concept-based semantic representation technique that maps text, for our problem, user query, into a high-dimensional space of Wikipedia concepts. Each dimension corresponds to a Wikipedia article, and the relevance of that concept to the input text is quantified using TF-IDF weighting based on the shared term frequency.

Let T be a query-term combination and $D = \{d_1, d_2, \ldots, d_N\}$ be the set of Wikipedia articles used as concepts, where each d_j corresponds to a distinct Wikipedia article. Let w_i be a word in T and v_T denotes the ESA vector representation of T. $\text{TFIDF}(w_i)$ represents the TF-IDF weight of word w_i and k_j be an inverted index entry for word w_i, that quantifies the strength of association between word w_i and Wikipedia article d_j, where $d_j \in D$.

$$v_T[j] = \sum_{w_i \in T} \text{TFIDF}(w_i) * (k_j) \qquad (1)$$

for each dimension $j = 1, \ldots, N$, where N is the number of Wikipedia concepts.

3.4 Top Query-Term Combination with k-Ranked Wikipedia Entities

We compute the cosine similarity between the ESA vector of each query-term combination and the corresponding matched Wikipedia entities. This step aims to identify a sequence of Wikipedia entities that best reflects various combinations of query terms. Based on extensive experimental analysis, we rank the top *k* entities, which are found to be optimal for our task, where $k = 3$. Based on the top-3 entities, we construct non-overlapping query-term combination, that represents query interpretation.

3.5 Document Retrieval

This query interpretation then is passed to an external search engine (in our case, Google), which retrieves relevant documents or search results. The search engine is independent and robust enough to evaluate the query interpretation and return results that closely align with the original query's intent.

4 Experiment Details

4.1 Data and Evaluation

We use the English Wikipedia dump[2] as our knowledge base, indexing article metadata including titles, full content, redirects, disambiguation flags, and internal page links, following standard practices. To ensure relevance, we exclude non-article content such as templates, user pages, user talk pages, Wikipedia management pages, file pages,

[2] https://dumps.wikimedia.org/

Question Type	Question Type	Baseline	CBQT	ESAQueryRank
“Where”	92	0.39	**0.47**	0.41
“How XX”	80	0.16	0.13	**0.17**
“When”	67	0.29	**0.42**	0.36
“Name a”	20	0.38	0.44	**0.53**
“Name the”	8	0.38	0.52	**0.59**
“How”	5	0.27	0.51	**0.63**
“Why”	4	0.05	0.11	**0.29**

Table 1: The detailed MRR comparisons with previous work on google search results. We use the paired t-test with significance at p $\leq$ 0.05.

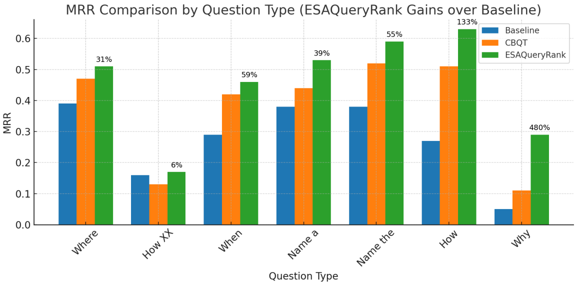

Figure 2: MRR comparison across question types. Percentage gains over the baseline and CBQT are annotated above the green bars.

category pages, and help documents. All data is stored and queried efficiently using SQLite.

We evaluate our open-domain question answering system using the TREC[3] question collection on a random sample of 270+ natural language queries. These are manually categorized into seven types, as shown in Table 1. The system feeds the query interpretation to external search engine Google for document retreival. For each original question, we retrieve the top $N = 10$ results and compare titles and snippets against TREC's gold-standard answers. We compare our approach against the baseline strategy and CBQT(Wang et al., 2024).

4.2 Experimental Results

Table 1 presents the MRR comparison between the baseline, Context-Based Query Translation from Wang et al. (2024), and our proposed approach, ESAQueryRank. As shown, ESAQueryRank outperforms the baseline in five out of seven question categories, achieving a maximum improvement of **480%** in the “Why” category. Our approach consistently demonstrates substantial improvements across almost all question types, with MRR increases exceeding **30%** over the baseline. This indicates ESAQueryRank's effectiveness in generating semantically aligned queries, especially in contexts where surface-level matching is insufficient. Figure 2 further illustrates the comparative performance of CBQT and ESAQueryRank against the baseline on Google search results. Notably, question types such as “How” and “Why” show significant gains, reflecting ESAQueryRank's ability to better interpret user intent and retrieve more relevant answers with greater precision than both the baseline and CBQT.

5 Conclusion

ESAQueryRank is an unsupervised, concept-based framework for interpreting and ranking entity and query-term combinations from natural language queries. Being based on ESA, it generates semantically coherent interpretations without query expansion, supervised learning based embeddings.

ESAQueryRank has been shown promosing results over keyword-based baseline and the Context-Based Question Translation model (CBQT) in particular for semantically complex query types like “Why” and “How.” Its transparent design and use of open-domain knowledge sourced from Wikipedia make it scalable and domain-independent.

The main contribution lies in a semantically rich query understanding model that avoids any query interpretations based on lexical overlap or dense embeddings but relies on Wikipedia-derived concept vectors without any defined training data or normalized concepts. This way, by grounding disambiguation and ranking in Wikipedia concept space, we increase the quality of query interpretations in ways that do not rely on labeled data. As a future work, We plan to combine ESA-based scores with the output of pre-trained language models (e.g., BERT) while keeping an interval notion of interpretability and evaluate on large-scale benchmarks such as SQuAD2.0.

[3] https://trec.nist.gov/data/qa.html

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, page 1606–1611, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Vishal Gupta and Ashutosh Dixit. 2023. Recent query reformulation approaches for information retrieval system-a survey. *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, 16(1):94–107.

Samreen Kazi, Shakeel Khoja, and Ali Daud. 2025. Bridging the gap: A survey of document retrieval techniques for high-resource and low-resource languages. *Computer Science Review*, 57:100756.

Minghan Li, Ming Li, Kun Xiong, and Jimmy Lin. 2021. Multi-task dense retrieval via model uncertainty fusion for open-domain question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 274–287, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dong Liu and Sreyashi Nag. 2025. Query brand entity linking in e-commerce search. *arXiv preprint arXiv:2502.01555*.

Jie Liu and Barzan Mozafari. 2024. Query rewriting via large language models. *arXiv preprint arXiv:2403.09060*.

Lajanugen et al. Logeswaran. 2019. Zero-shot entity linking by reading entity descriptions. In *ACL*.

Gerard Pons, Besim Bilalli, and Anna Queralt. 2024. Knowledge graphs for enhancing large language models in entity disambiguation. In *International Semantic Web Conference*, pages 162–179. Springer.

Martin Potthast, Benno Stein, and Maik Anderka. 2008. A wikipedia-based multilingual retrieval model. In *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval*, ECIR'08, page 522–530, Berlin, Heidelberg. Springer-Verlag.

Tsegaye Misikir Tashu, Eduard-Raul Kontos, Matthia Sabatelli, and Matias Valdenegro-Toro. 2025. Cross-lingual document recommendations with transformer-based representations: Evaluating multilingual models and mapping techniques. In *Proceedings of the Second Workshop on Scaling Up Multilingual & Multi-Cultural Evaluation*, pages 39–47.

Haofen Wang, S. M. Mazharul Hoque Chowdhury, and Wei Jin. 2024. Wikipedia empowered natural language interface for web search. In *Web Information Systems Engineering – WISE 2024: 25th International Conference, Doha, Qatar, December 2–5, 2024, Proceedings, Part I*, page 14–25, Berlin, Heidelberg. Springer-Verlag.

Ledell et al. Wu. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *EMNLP*.

Faren Yan, Peng Yu, and Xin Chen. 2025. Ltner: Large language model tagging for named entity recognition with contextualized entity marking. In *International Conference on Pattern Recognition*, pages 399–411. Springer.

Zi Ye, Yogan Jaya Kumar, Goh Ong Sing, Fengyan Song, and Junsong Wang. 2022. A comprehensive survey of graph neural networks for knowledge graphs. *IEEE Access*, 10:75729–75741.