In this field, Inner Cover information will be filled for the thesis.

Fill in this field in accordance with the information specified in the Thesis Writing Guide so that it does not overflow the physical outer cover frame.

PURELY ENTITY-BASED SEMANTIC SEARCH FOR INFORMATION RETRIEVAL

Mohamed Lemine SİDİ

PhD Dissertation

Department of Computer Engineering

Programme in Computer Science

Supervisor: Prof. Dr. Serkan GÜNAL

Eskişehir

Eskişehir Technical University

Institute of Graduate Programs

July 2023

FINAL APPROVAL FOR THESIS

This thesis titled PURELY ENTITY-BASED SEMANTIC SEARCH FOR INFORMATION RETRIEVAL has been prepared and submitted by Mohamed Lemine SİDİ in partial fulfillment of the requirements in "Eskişehir Technical University Directive on Graduate Education and Examination" for the Degree of PhD in Computer Engineering Department has been examined and approved on 26/07/2023.

Committee Members		**Title, Name and Surname**	**Signature**
Member	:	Title Name SURNAME	
Member	:	Title Name SURNAME	
Member	:	Title Name SURNAME	
Member	:	Title Name SURNAME	
Member	:	Title Name SURNAME	

Prof. Dr. Semra KURAMA
Director of the Institute of Graduate Programs

26/07/2023

SUPERVISOR APPROVAL

PhD student Mohamed Lemine SİDİ, whom I supervise, has completed this thesis titled PURELY ENTITY-BASED SEMANTIC SEARCH FOR INFORMATION RETRIEVAL. According to my inspections, the work is scientifically and ethically appropriate for the student to the thesis defense exam.

Supervisor
Prof. Dr. Serkan GÜNAL

ABSTRACT

PURELY ENTITY-BASED SEMANTIC SEARCH FOR INFORMATION RETRIEVAL

Mohamed Lemine SİDİ

Department of Computer Engineering

Programme in Computer Science

Eskişehir Technical University, Institute of Graduate Programs, July 2023

Supervisor: Prof. Dr. Serkan GÜNAL

Over the past decade, Knowledge bases (KB) have been increasingly used in almost all information retrieval tasks. To improve the ad hoc document retrieval task, KB have often been utilized to complete and enrich the representation of queries and documents. Although many approaches have used KB for such purpose, understanding how effectively leverage entity-based representation in conjunction with term-based representation still needs to be resolved. In this thesis, we propose a Purely Entity-based Semantic Search Approach for Information Retrieval (PESS4IR). We explore its strengths and weaknesses to know how it would be effectively leverageable alongside any other model. The approach includes (i) its own entity linking method, Entity Linking for Document Text (EL4DT), which is designed to be appropriate for document text. Moreover, the approach includes (ii) an inverted indexing method for the indexing task. For document retrieval and ranking, (iii) an appropriate ranking method is designed to take advantage of all the strengths of the approach. We report the findings on the performance of our approach tested by queries annotated by two entity linking tools, REL and DBpedia Spotlight. The experiments are performed on the standard TREC 2004 Robust collection and MSMARCO collections. By using the REL method on Robust collection, for queries whose all terms are annotated and whose average annotation scores are greater than or equal to 0.75, our approach achieves the maximum nDCG@5 score (1.000). Thus, using our approach with any document retrieval method would be an added value, unless that method achieves the maximum nDCG@5 score for those highly annotated queries.

Keywords: Information Retrieval, Knowledge graph, Entity-based Search, Document Retrieval, Entity Linking.

ÖZET

BİLGİ ERİŞİMİ İÇİN TAMAMEN VARLIK TABANLI SEMANTİK ARAMA

Mohamed Lemine SİDİ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Bilimleri Bilim Dalı

Eskişehir Teknik Üniversitesi, Lisansüstü Eğitim Enstitüsü, Temmuz 2023

Danışman: Prof. Dr. Serkan GÜNAL

Son on yılda, Bilgi tabanları (Knowledge bases) neredeyse tüm Bilgi Erişimi görevlerinde giderek daha fazla kullanılmaktadır. Ad-hoc belge erişimi görevini iyileştirmek için, sorguların ve belgelerin sunumunu tamamlamak ve zenginleştirmek için genellikle KB kullanılmıştır. Pek çok yaklaşım KB'yi bu amaçla kullanmış olsa da, terim tabanlı temsil ile birlikte varlık tabanlı temsilden ne kadar etkili bir şekilde yararlanıldığının anlaşılmasının hala çözülmesi gerekmektedir. Bu tezde, Bilgi Erişimi için Tamamen Varlık Tabanlı Semantik Arama Yaklaşımı (PESS4IR) önerilmiştir. Diğer herhangi bir modelin yanında nasıl etkili bir şekilde kullanılabileceğini bilmek için güçlü ve zayıf yönlerini keşfedilmiştir. Yaklaşım, (i) belge metnine uygun olacak şekilde tasarlanmış kendi varlık bağlama yöntemini, Belge Metni için Varlık Bağlama'yı (EL4DT) içerir. Ayrıca, yaklaşım (ii) indeksleme görevi için tersine çevrilmiş bir indeksleme yöntemini içerir. Belge alma ve sıralama için, (iii) yaklaşımın tüm güçlü yönlerinden yararlanmak için uygun bir sıralama yöntemi tasarlanmıştır. İki varlık bağlama aracı, REL ve DBpedia Spotlight tarafından açıklamalı sorgularla test edilen yaklaşımımızın performansına ilişkin bulguları rapor edilmiştir. Deneyler, standart TREC 2004 Robust koleksiyonu ve MSMARCO koleksiyonları üzerinde gerçekleştirilmiştir. Robust koleksiyonunda REL yöntemini kullanarak, tüm terimleri açıklamalı ve ortalama açıklama puanları 0,75'ten büyük veya buna eşit olan sorgular için yaklaşımımız maksimum nDCG@5 puanını (1.000) elde eder. Bu nedenle, yaklaşımımızı herhangi bir belge alma yöntemiyle kullanmak, bu yöntem yüksek düzeyde açıklamalı sorgular için maksimum nDCG@5 puanına ulaşmadığı sürece, katma bir değer olacaktır.

Anahtar Sözcükler: Bilgi Erişimi, Bilgi Grafiği, Varlık-dayalı Arama, Belge Erişimi, Varlık Bağlama.

ACKNOWLEDGEMENTS

This thesis is the final result of great efforts, at the head of which comes the efforts of my parents in raising and educating me in an atmosphere of love and tenderness. This thesis is dedicated to them, my beloved mother, and to my father's soul.

I would first like to thank my supervisor, Prof. Dr. Serkan GÜNAL, for his guidance, advice, motivation, and positive spirit during difficult times. I also thank him for his excellent academic treatment and high morals, as he never addressed me except with elegant words. I wish to thank the rest of my Ph.D. committee for their advices throughout the follow-up period, Assoc. Prof. Dr. Ahmet ARSLAN and Assist. Prof. Dr. Zuhal KARTAL. Moreover, I would like to thank the teaching staff of my department, especially those from whom I took courses.

I would like to extend my thanks to my friends and colleagues, especially Celal AYDOĞDU, Mohamed El Benani EKHLIVA, Mehmet ÖZCAN, Abdulatif Hajaismael Ahmed, Mahmoud Ahmad, and Aman Sado who have helped me in numerous ways over the years.

I convey my heartfelt thanks to each member of my family and friends who have been sending me their love and prayers over the years.

Mohamed Lemine SİDİ

26/07/2023

STATEMENT OF COMPLIANCE WITH ETHICAL PRINCIPLES AND RULES

I hereby truthfully declare that this thesis is an original work prepared by me; that I have behaved in accordance with the scientific ethical principles and rules throughout the stages of preparation, data collection, analysis and presentation of my work; that I have cited the sources of all the data and information that could be obtained within the scope of this study, and included these sources in the references section; and that this study has been scanned for plagiarism with "scientific plagiarism detection program" used by Eskişehir Technical University, and that "it does not have any plagiarism" whatsoever. I also declare that, if a case contrary to my declaration is detected in my work at any time, I hereby express my consent to all the ethical and legal consequences that are involved.

Mohamed Lemine SİDİ

CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

GLOSSARY OF SYMBOLS AND ABBREVIATIONS

α	:	Alpha
β	:	Beta
APA	:	American Psychological Association
QF-EHEA	:	Qualifications Framework for European Higher Education Area
SI	:	The International System of Units
	:	
	:	

1. INTRODUCTION

Document retrieval also called ad-hoc retrieval is the core task of information retrieval (IR). The purpose of the ad-hoc retrieval task is to find the relevant documents that satisfy the information needs textually expressed in the query. Search engines are the simplest example of it. They provide a ranked list of relevant documents (web pages), for a given query. Moreover, over the last decade, semantics-based approaches have become increasingly investigated and used for ad-hoc retrieval task, due to their significant contributions.

1.1. Document Retrieval with Knowledge Bases

Because of their semi-structured rich, and strong semantics, knowledge bases (also called knowledge graphs) are exponentially used in ad-hoc retrieval task and generally in the different information retrieval tasks. Furthermore, the quality and quantity of the knowledge bases such as DBpedia (Lehmann, 2015) are continuously increasing, which gives us an idea of the knowledge bases' usefulness now and in the future (Dietz, 2018). Moreover, knowledge bases such as DBpedia and Freebase (Bollacker, 2008) are among the most widely used ones.

Based on knowledge bases, a given text could be represented by a suitable set of entities. This representation of text by entities would be called entity-based representation. Moreover, there are several ways to utilize knowledge bases to improve the representation of queries and documents for better ad hoc document retrieval task. When entity-based representation is used alongside term-based representation for query representation, in this case, knowledge bases are used for query expansion ((Brandão, 2014), (Dalton, 2014)). Furthermore, it is used for both query and document representation completing and enriching the term representation for better document retrieval ((Xiong C. C., Word-entity duet representations for document ranking, 2017), (Liu, 2018), (Lashkari, 2019)).

Although many entity-based representation approaches have been proposed for semantic document retrieval, understanding how effectively to leverage entity-based representation still needs to be resolved ((Reinanda, 2020), (Guo, 2020)). To answer the question "How to know when an approach (a retrieval method) does its best, and for what type of queries?" it is essential to explore the strengths and weaknesses of the approach and, on the other side, analyze its performance towards different queries. On the other hand, because of the complexity and dynamic nature of queries, no document retrieval approach achieves the same performance for every query. Although a better approach has a higher performance score regarding the whole query set, a lower-quality approach may have better performance for some of those queries. Once these questions are effectively addressed, a document retrieval system, such as a search engine, could leverage many approaches alongside each other regarding the different kinds of queries for better information retrieval.

We propose a novel semantic search approach, named Purely Entity-based Semantic Search for Information Retrieval (PESS4IR), purely based on entity representation, and explore its strengths and weaknesses for better document retrieval. In other words, the "Purely entity-based representation" concept means that documents and queries are only represented by entities. The approach is mainly composed of three components. The first one is its own entity linking method, which is more appropriate for document text, named Entity Linking for Document Text (EL4DT). We note that, for the entity linking task, there are many tools available online, such as DBpedia Spotlight (Mendes, Jakob, García-Silva, & Bizer, 2011), TagMe (Ferragina & Scaiella, 2010), and REL (van, Hasibi, Dercksen, Balog, & de Vries, 2020). However, the main reason for developing a new entity linking (EL4DT) method is that these tools do not provide certain information and statistics necessary for our document retrieving and ranking method. The second component is an inverted indexing method to achieve the indexing task. The third and last one is an appropriate document retrieving and ranking strategy leveraging all strengths of the approach.

A visual representation of PESS4IR is provided in Figure (1.1), which explains the offline and online time of the approach. The offline time provides the preprocessing method, which cleans and splits document text to be prepared for the entity linking task, achieved by our entity linking method EL4DT. Once the documents are annotated, their entity-based representation is indexed by our inverted indexing method. On the other hand, the online time represents the inference time, which contains the retrieval and ranking tasks carried out by our ranking method.

Our approach introduces the "strong entity" concept to describe entities annotated by EL4DT with high scores. The concept plays an essential role in our retrieval and ranking method. Moreover, our approach introduces the "related entities number" concept to describe the number of entities semantically related to an entity, for each entity in the paragraph. Furthermore, alongside the "strong entity" and "related entities number" concepts, our retrieval and ranking method leverages many other aspects of our approach, such as document title weighting and all information and statistics stored in the index, such as the number of semantically related entities in the same paragraph (see Indexing session).

Before discussing the evaluation of our approach, it is vital to understand the nature of the purely entity-based representation. In fact, queries are ambiguous (Xiong C. C., 2017) due to the nature of their text, which is usually short and suffers from a lack of context, contrary to document text, which could be well represented by entities because of its textual richness. Annotating the query set of the TREC 2004 Robust collection (250 queries) by DBpedia Spotlight and REL entity linking tools highlights the completely annotated query concept (all query terms are annotated, stopwords are not necessarily annotated). Whereas DBpedia Spotlight and REL annotate 72% and 6.8% of queries, respectively, as completely annotated queries (see appendices A.1 and A.2). Thus, in our experiments, only completely annotated queries are considered. Since our approach is designed to handle only completely annotated queries, the evaluation will be on the corresponding partial results. Therefore, in the evaluation, it is necessary to consider the partial nature of the results. Moreover, for our experiments, we use TREC 2004 Robust and MSMARCO collections. Furthermore, we use the well-known baseline Galago tool[1], a search engine extended by the Lemur and Indri projects for research purposes (Croft, 2010), to get the corresponding partial results performed by Galago (Dirichlet method) for all our experiments, on Robust collection, to show the added value of our approach. Also, alongside Galago, LongP (Longformer) (Beltagy, 2020) model is used for comparisons. We use REL and DBpedia Spotlight entity linking tools for the query annotation process as arbitrary entity linking methods instead of our entity linking method. In other words, our approach is tested with the queries annotated by other methods. Also, in the evaluation, we use standard evaluation metrics such as nDCG@k, MAP, and P@k.

In the thesis, we explore the strengths and weaknesses of our approach by taking advantage of its strengths and avoiding its weaknesses. The exploration study allowed us to answer the question for which query a purely entity-based approach is recommended. Furthermore, for queries with an average annotation score (average annotation score of

[1] https://sourceforge.net/projects/lemur/files/lemur/galago-3.15/

query entities) higher than 0.75, annotated by the REL method, our approach achieves the maximum nDCG@5 score (1.000), which would be an added value for any ad-hoc document retrieval method which does not reach the maximum nDCG@5 score achieved on the Robust collection.

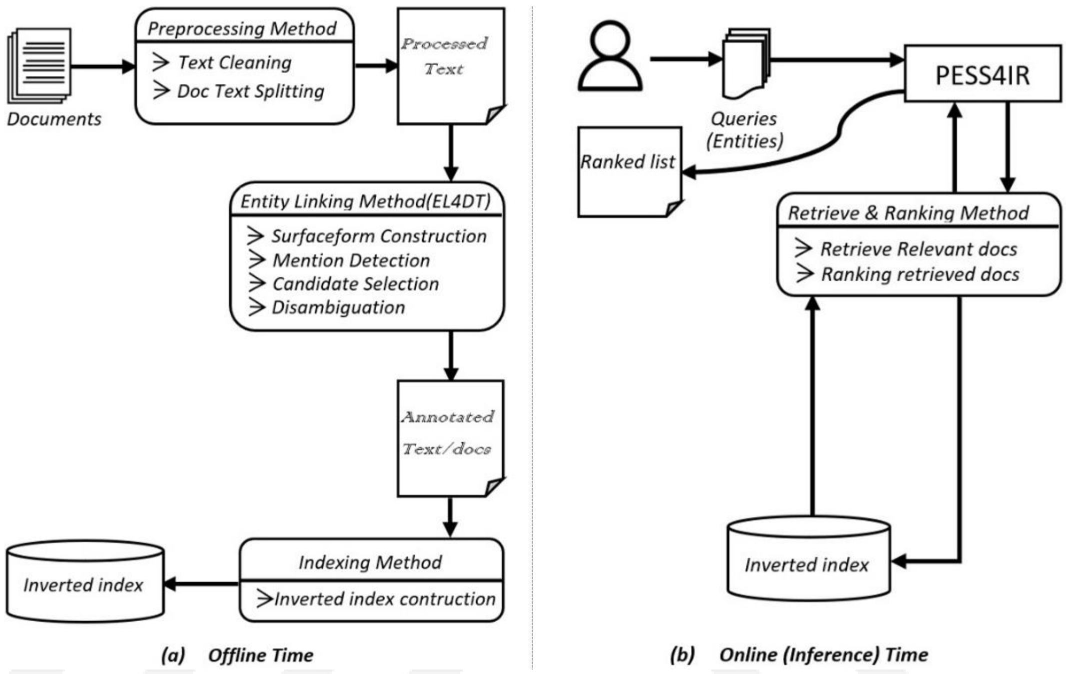

Figure 1.1. *The visual representation of our approach.*

1.2. Contributions of Thesis

The main contributions of this dissertation are as follows:

We introduce PESS4IR, a novel approach designed and developed for better ad-hoc document retrieval, and purely based on knowledge graphs. The approach includes:

- Designing and developing an entity linking method appropriate to document text (EL4DT) and our solution. The method provides specific information and statistics needed by our approach.
- Developing an inverted indexing method to achieve the indexing task suitable to our solution.
- Designing and developing a retrieval and ranking method appropriate to our approach.

PESS4IR gives answers about when a purely entity-based semantic approach does its best and when it does not, showing its strengths and weaknesses, to leverage strengths and then use it when it is recommended to be used.

Thus, the three elements of our approach mentioned above have been regrouped in one published article (Sidi & Gunal, 2023).

1.3. Organization of Thesis

This dissertation is organized as follows. Chapter 2 provides the background and the related works, presenting some state-of-the-art models for the document retrieval task. And then focusing on similar models, which use knowledge bases, in Session 2.1. Also provides, in Session 2.2, the main related works of Entity Linking, which is a related task to the knowledge bases-based ad-hoc retrieval task. Finally, it presents another secondary concern, the query performance prediction task, in Session 2.3. Chapter 3 represents the whole approach including the Entity linking method (EL4DT), the indexing method, and the retrieval and ranking method. Chapter 4 provides the achieved experiments and their results, in Session 4.4. Session 4.5 provides a discussion of the experiments and their results. Finally, Chapter 5 presents the conclusion highlighting the strengths and weaknesses of the approach, and finishes by providing some perspectives.

2. BACKGROUND AND RELATED WORKS

In this Chapter, we present an overview of the ad-hoc retrieval task and the entity-based document retrieval, our primary concerns in this thesis, and the entity linking and query performance prediction tasks as secondary concerns.

2.1. Document Retrieval Task

Semantic search significantly improves information retrieval (IR) tasks, including ad hoc document retrieval task ((Xiong C. C., 2017), (Liu, 2018), (Lashkari, 2019), (Guo, 2020), (Xiong C. C., 2016), (Dehghani, 2017), (Bagheri, 2018), (Zamani, 2018), (Xiong C. D., 2017)). Moreover, knowledge bases have been used increasingly used to improve the semantic search ((Xiong C. C., 2017), (Lashkari, 2019), (Xiong C. C., 2016), (Bagheri, 2018), (Liu, 2018)). Knowledge bases allow an entity-based representation of text instead of the original words used in lexical representation and traditional models such as the BM25 model (Robertson, 2009). Moreover, in this section, we present some state-of-the-art document retrieval models which are not based on knowledge bases.

The approaches based on pre-trained Transformer language models such as BERT (Devlin, Chang, Lee, & Toutanova, 2018) are the current state-of-the-art for text re-ranking (Gao & Callan, 2022). Such a neural language model learns the contextual features of words, where the representation of a word is a function of the entire input text, with word dependencies and sentence structures taken into consideration (Dai & & Callan, 2019).

In (Li, Yates, MacAvaney, He, & Sun, 2020) research, they proposed a a neural language model named PARADE, which is a re-ranking model. They claimed an improvement achieved by the model on TREC Robust04 and GOV2 collections; where the model achieves its most effective performance when it is adopted as a re-ranker namely PARADE-Transformer. The Longformer model (Beltagy, 2020) is also a neural language model. Its performance on Robust04 and MSMARCO collections is provided in (Boytsov, 2022). We use the LongP Longformer model (Boytsov, 2022) in our comparisons against state-of-the-art methods. Moreover, in (Gao & Callan, 2022), they proposed the MORES+, a neural language model, which is a re-ranking method, and tested it on two classical IR collections Robust04 and ClueWeb09. In (Wang, Macdonald, Tonellotto, & Ounis, 2023) work, they proposed a ranking method named ColBERT-PRF. They evaluated it on MSMARCO document ranking and TREC Robust04 for document ranking tasks. The approach could be exploited for both end-to-end ranking and reranking scenarios.

2.2. Entity-based Document Retrieval

In the literature, over the last decade, many document retrieval approaches have explored different types of using entity-based representation for improving the representation of documents and queries. In (Xiong C. C., 2017) work, they proposed a neural information retrieval approach using both term-based and entity-based representations for queries and documents. The approach performs the four-way interactions allowing four matching possibilities between query and document. The four possible representations of document and query are query terms (words) to document terms, query entities to document entities, query entities to document terms, and query terms to document entities. Thus, they achieve their document retrieval and ranking by integrating those combinations into neural models. In (Liu, 2018), they also, introduced a neural ranking model that combines entity-based and term-based representations of documents and queries. They used a translation layer in their neural architecture, matching queries and documents without using handcrafted features. In (Bagheri, 2018), a document retrieval approach was proposed, which uses neural embeddings considering both word embeddings and entity embeddings. They compared both word and entity embedding performances. (Lashkari, 2019) proposed a neural embeddings-based representation for documents by considering the term, entity, and semantic type within the same embedding space. (Gerritse, 2022) proposed EM-BERT model, which incorporates entity embeddings into a point-wise document ranking approach. The model combines words and entities into an embedding representation to represent both query and document using the BERT model (Devlin, Chang, Lee, & Toutanova, 2018).

Although many entity-based document retrieval approaches have been proposed, understanding how effectively to leverage entity-based representation still needs to be resolved. (Guo, 2020) presented a survey on existing neural ranking models, highlighting models that learn with external knowledge, such as knowledge bases. They indicated that more research is needed to improve the effectiveness of neural ranking models with external knowledge and understand external knowledge’s role in ranking tasks. Moreover, (Reinanda, 2020) describes the fact that understanding how effectively leverage entity-based representation in conjunction with term-based representation still needs to be solved. On the other hand, these approaches and models use knowledge graphs as embedding-based representations, where entity embeddings are learned from knowledge graphs in many ways in the literature ((Bordes, 2013), (Lin, 2015), (Schuhmacher, 2015)). Although the neural ranking models play an important role in the document retrieval task, they have also their own issues and weaknesses. In (Wu C. Z., 2022), they investigated the robustness of neural ranking models and came to the conclusion that neural ranking models are less robust compared to other IR models in most cases.

To understand how effectively leverage knowledge graph alongside any other model, we introduced PESS4IR, a novel solution, to empirically study the impact of purely entity-based representation for document retrieval. With PESS4IR, queries and documents are represented only by an entity-based representation without using a neural network and embedding-based representation. Below, we discuss the background related to the entity-based document retrieval task.

2.3. Entity Linking

Entity linking is the task that links terms of a given text to appropriate entities extracted from Knowledge bases; in other words, it gives an entity-based representation that is suitable for the given text. There are many entity linking tools, among which DBpedia Spotlight (Mendes P. N.-S., 2011), TagMe (Ferragina P. &., 2010), REL (van H. M., 2020), WAT (Piccinno, 2014), and FEL (Pappu, 2017) are the most widely used. Most of these entity-linking tools are designed for general text annotation purposes. Moreover, some of them would perform better for short text than others, such as TagMe, which is known for its performance for short text (Chen, 2018). However, to get a more appropriate entity linker for document text, we develop a novel entity linking method, which gives specific information and statistics our approach needs. In the literature, an effective entity linking method is generally designed based on the general pipeline, which is composed of three steps: mention detection, candidate selection, and disambiguation. Moreover, disambiguation is the most challenging step (Balog, 2018) in the entity linking task. According to (Balog, 2018), modern disambiguation should consider three important types of evidence: prior importance, contextual similarity, and coherence. Many researchers deal with disambiguation via a graph-based approach. Recently (Kwon, 2021), they dealt with the disambiguation issue by proposing a graph-based solution. Also, our entity linking method (EL4DT) uses a graph-based method for the disambiguation task, and the three types of evidence described in (Balog, 2018) are considered.

2.4. Query Performance Prediction

As our approach uses a purely entity-based representation for both query and document, it deals only with completely annotated queries. Moreover, it reaches its best performance when the given query (among completely annotated queries) has a high annotation score; thus, the higher the query annotation score, the better the performance of the approach. In other words, the query performance depends on the query entities' annotation scores. In the literature, this task is called query performance prediction (QPP). QPP has two types of retrieval predictors: pre-retrieval (Zendel, 2021) and post-retrieval (Datta, 2022). In (Zendel, 2021), they describe the post-retrieval predictors as more accurate than pre-retrieval predictors but usually more expensive to compute. However, with our approach, without developing a QPP method, the query performance is predicted by the pre-computed information, which is the query annotation average score. However, it is important to note that even the entity linking used method for query annotations it is possible to have some wrong annotations. We discuss the positive and negative sides of the prediction of query performance prediction in more detail in the discussion session.

2.5. Summary

This chapter first reviews the document retrieval task, providing the state-of-the-art of the task. Then it provides the recent works of document retrieval, which are based on knowledge graphs. After that, it provides an overview of entity linking tools as a related field. The last session presents some related works about the query performance prediction task and how it would be related to our approach, as a secondary concern.

3. PROPOSED APPROACH

This Chapter presents our approach (PESS4IR), which mainly includes three methods: entity linking method (EL4DT), indexing method, and retrieval and ranking method. In other words, PESS4IR starts with the text preprocessing task, which prepares documents to be annotated by our entity linking method. Then, the indexing method takes place to index the output, annotated documents, of the entity linking method. Finally, the retrieval method retrieves and ranks documents from the index based on the annotations of a given query. In the following sessions, we explain each technique in detail.

3.1. Preprocessing

The disambiguation algorithm of EL4DT is based on the paragraph's main idea and document coherence aspects. It is how our entity linking method is designed. We split the document text into paragraphs in the text preprocessing step to make our entity linking method easier. In addition, before the document splitting, the text cleaning task is performed, and stopwords are removed.

3.2. Entity Linking Method

We needed to design and develop an entity-linking method for document annotation, knowing that many available entity-linking methods exist. The main reason is to provide our approach with some required information and statistics, which are not provided by the available entity linking tools. We store the information and statistics in the inverted index to make them available for our retrieval and ranking method. In the following subsections, we explain the steps of our entity linking method.

3.2.1. Overview

Our entity linking method is designed precisely for the document text, named Entity Linking for Document Text (EL4DT). It is based on two knowledge bases, DBpedia (Lehmann, 2015) and Facc1 (Gabrilovich, 2013), from which the surface forms are constructed. In addition, EL4DT respects the general pipeline (Figure 3.1), which supposes that an entity linking method considers three steps: mention detection, candidate selection, and disambiguation. They are explained in more detail in the following three subsections, respectively.

Figure 3.1. *General pipeline of an Entity linking method.*

Before going into the details, we briefly describe our Entity Linking method by explaining its three main steps. In the first step, which contains mention detection and candidate selection tasks, the process starts with a given preprocessed text, and by the n-gram method, mentions are generated. For each generated mention, the candidate entities are extracted from the surface form. During the first step, a pre-disambiguation process is performed by selecting, in the case of many entities proposed by one class of the surface form classes Table (3.2), the more probable candidate entity according to the context similarity score. Moreover, for each candidate entity, the score is computed. In the second step, we have the disambiguation method, which performs the disambiguation task using graphs. Figure (3.2) represents an initialized graph for entities of a given paragraph, where each entity (represented by a node in the graph) could be connected to another if there are some relationships between them. Their relationships (defined by an edge in the graph) are scored according to the nature of those relationships. Besides, the entities which are semantically related will be grouped as a cluster of the graph. Also, the relationships between entities are found in article categories (DBpedia), SKOS's relationships, such as broader and related (DBpedia), and document coherence relationships. In the last step, we select the graph's highest score cluster. The selected graph cluster, a set of entities among the paragraph's entities, includes the disambiguated entities. It is noted that sure entities (with a score greater than or equal to 0.5) do not need a disambiguation step. Only weak entities which are not related to the selected cluster are ignored. Moreover, the frequently used symbols are listed in Table 3.1.

Table 3.1. *Used symbols.*

Symbol	Description
E	All entities.
E_q	Entities in query q.
E_p	Entities in paragraph p.
E_{qp}	Common entities between query q and paragraph p.
E_{dt}	Entities in document title dt.
SE_p	Strong entities in paragraph p.
SE_d	Strong entities in document d.
SE_{qp}	Entities in query q, which are found in paragraph p as strong entities.
T_d	Document text
T_p	Paragraph text

Furthermore, Algorithm (1) contains the disambiguation algorithm as a part of the EL4DT algorithm.

3.2.2. Mention Detection

A mention refers to a contiguous sequence of terms in the text to be annotated, which refers to one or more particular entities in the Surface form (Balog, 2018). The surface form is the structure that includes all possible mentions extracted from knowledge bases. As mentioned earlier, our surface form is based on DBpedia and Facc1 knowledge bases. The surface form of our EL4DT is constructed from the components listed in Table 3.2.

Table 3.2. *Components of the Surface form.*

Class	Description	Knowledge Base
E_db	Entities extracted from Article categories (without stopwords)	DBpedia
cE_db	Entities extracted from Article categories (with stopwords)	DBpedia
ED_db	Entities extracted from Disambiguation (without stopwords)	DBpedia
cED_db	Entities extracted from Disambiguation (with stopwords)	DBpedia
RE_db	Entities extracted from Redirects (without stopwords)	DBpedia
cRE_db	Entities extracted from Redirects (with stopwords)	DBpedia
E_dbFacc	Common entities extracted from FACC1 and DBpedia's Article categories (without stopwords)	Facc1 and DBpedia
cE_dbFacc	Common entities extracted from FACC1 and DBpedia's Article categories (with stopwords)	Facc1 and DBpedia
E_Similar	Upper- and lower-case modified entities from DBpedia's Article categories	DBpedia

For a given text, which is supposed to be a paragraph, an n-gram method (n = 8) is used to find all possible candidate entities corresponding to each mention. The candidate entities are extracted from the surface form. We note that the (n = 8) is chosen according to the number of terms in entities from the surface form, where for n greater than eight terms, entities in Facc1 and DBpedia knowledge bases are not much. Thus, we ignore the small number of entities whose number of terms is larger than eight. Therefore, from a given paragraph, all possible mentions, which exist on the surface, would be detected.

3.2.3. Candidate Selection

There is at least one candidate entity for each mention. Moreover, some candidate entities could be included in others, where the role of the candidate selection algorithm is to select the more probable candidate entity for each mention. For example, if we consider the following sequence of words "The Empire State Building …"; with the mention detection step, the three following mentions could be detected from the surface form: "Empire", "Empire State", and "Empire State Building". One can observe that the first two mentions are included in the third one; if there are one or more candidate entities, in the surface form, for the third mention, then the first and second the mention will be ignored. In the same way, candidate entities detected for the first and second mentions will be ignored.

The selection score of a candidate entity is computed by considering different factors such as:

- The component weight: It is a defined weight according to each component of the surface form components (Table 3.2).
- Contextual similarity computations score: It is defined as a score of similarity between entity terms and the given paragraph.
- The number of terms in the entity.

Thus, these score computations are used in the candidate selection algorithm to select the most appropriate entity for each mention.

3.2.4. Disambiguation

The disambiguation task is achieved using a graph-based algorithm, which is the central part of our EL4DT method. The constructed graph is a weighted graph G = (V, E), where the node set V contains all selected candidate entities from a given paragraph, and each edge represents the semantic relationship between two entities. The main goal of the disambiguation algorithm is to select among ambiguated entities (entities with weak scores) only those related to sure entities (entities with scores higher than or equal to 0.5). Thus, other weak entities are ignored. Furthermore, EL4DT identifies the best cluster, the group of related entities with the higher score among other clusters in the graph. In other words, the best cluster in the graph is supposed to be the paragraph's main idea. In a graph, we note that a cluster is a set of entities connected by edges whose weights are greater than zero.

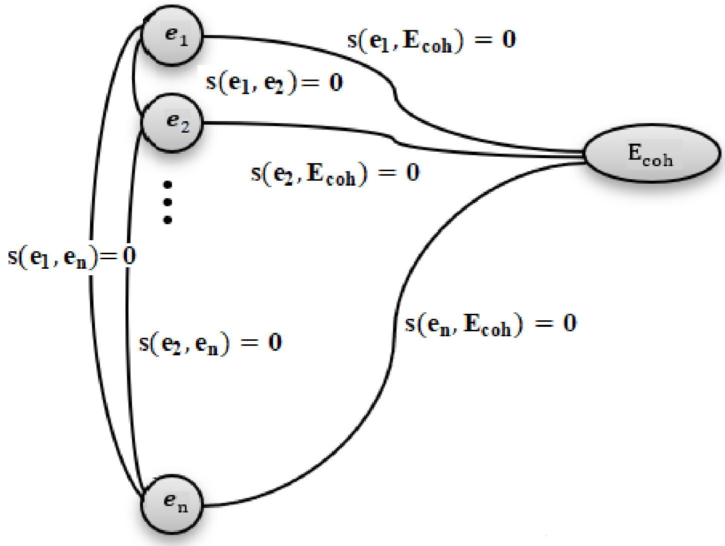

Figure 3.2. *Initialization of a graph built from annotated entities of a paragraph.*

The set of entities in a paragraph is expressed by $E_p = \{e_1, e_2, \ldots, e_n\}$, where n represents the number of entities in the paragraph.

The $\boldsymbol{E_{coh}}$ symbol stands for coherence entities, which exist in the document title and the document's strong entities. The strong entity concept refers to entities annotated by EL4DT with an annotation score higher than or equal to 0.85. Thus, $\boldsymbol{E_{coh}}$ represents the intersection between the document's strong entities $\boldsymbol{SE_d}$ and document title entities $\boldsymbol{E_{dt}}$, as Equation (3.1) shows.

$$E_{coh} = E_{dt} \cap SE_d \tag{3.1}$$

After the graph initialization step, the graph scoring expression is based on the Equation (3.2), which shows how graph edges are scored:

$$GS(e_i, e_j) = \frac{rScore(e_i, e_j) \times (LScore(e_i) + LScore(e_j))}{|E_p|}, \tag{3.2}$$

where $\mathbf{rScore}(\boldsymbol{e_i}, \boldsymbol{e_j})$ is the number of relationships between two entities, computed by (3.3), where R represents different types of relationships between two entities; moreover, the relationship is either direct or indirect. Direct relationships exist when entities share common article categories of DBpedia. Moreover, from the SKOS of DBpedia, whether there are some relationships between the entities such as <skos:broader> and <skos:related> predicates. However, an indirect relationship means that $\boldsymbol{e_i}$ and $\boldsymbol{e_j}$ have no direct relationships but rather are related by $\boldsymbol{E_{coh}}$ as an indirect relationship. The following Equations (3.3, 3.4, 3.5, and 3.6) explain Equation (3.2) respectively:

$$rScore(e_i, e_j) = n(e_i, e_j, R) \tag{3.3}$$

$$n(e_i, e_j, R) = \sum_{c \in ACatg(e_i)} \{exist(c, ACatg(e_j))\} + exist(e_i, SkosR(e_j)) + exist(e_j, SkosR(e_i)) + exist(e_i, e_j, E_{coh}) \tag{3.4}$$

$$exist(c, ACatg(e_j)) = \begin{cases} 1, c \in ACatg(e_j) \\ 0, \ otherwise. \end{cases} \tag{3.5}$$

$$exist(e_i, e_j, E_{coh})) = \begin{cases} 1, (e_i e_j) \in E_{coh} \\ 0, \ otherwise. \end{cases} \tag{3.6}$$

Moreover, the $\mathbf{LScore}(\boldsymbol{e_i})$ gives the entity score computed by EL4DT in the candidate selection process, which mainly calculates the three factors (assignment of components weights, score of contextual similarity, and number of terms in the entity) when considering the different surface forms and the three types of entity, such as the non-sure, sure, and strong entities. $ACatg(e_i)$ provides the entity category of the entity e_i, which includes all entities of the same DBpedia category. $SkosR(e_i)$ provides the Skos relations (and predicates) of the entity e_i, which are extracted from DBpedia.

3.2.5. EL4DT Algorithm

The following algorithm (Algorithm 1) represents an overview of our entity linking method. It gives the main processes of entity linking for a given document, starting with the document text (T_d) as input and finishing with the corresponding annotations ($document_{annotations}$) as output. Between line number 3 and 6, we have a loop, which represents the process of candidate entity selection. In that process, for each paragraph text (T_p) all possible candidate entities are detected. Moreover, an n-gram method is used to detect the candidate entities for a given text. Once all possible candidate entities are detected, the more probable ones are selected, for each mention. The rest of the algorithm shows how the disambiguation method works for document text. It starts with the graph initialization. Moreover, for each paragraph in the document, a new graph is built. Between lines (10-12), the graph scoring process is performed, with the help of Equation (3.2). Then, the disambiguation process is performed by the selection of disambiguated entity set in the graph, according to the cluster with the highest score in the graph and sure entities in other clusters. In the disambiguation subsection, we explain how the disambiguation method works. Finally, in line (14), the selected entity set is added to the annotations of the given document.

ALGORITHM 1: EL4DT algorithm (Mention Detection, Candidate Selection, Disambiguation)

Input: $T_d \leftarrow$ document_text
Output: $document\_annotations$
for $T_p \in T_d$ **do**
    ms $\leftarrow$ find_allPossible_candidate_entities (T_p)
    $E_p \leftarrow$ select_candidate_entity(ms, T_p)
end for
$E_{coh} \leftarrow E_{dt} \cap SE_d$
for $E_p \in E_d$ **do**
    G (V, E) $\leftarrow$ graph_initialization(E_p)
    for v, e $\in$ G **do**
        $e \leftarrow \frac{\mathrm{rScore}(e_i,e_j)\times(\mathrm{LScore}(e_i)+\mathrm{LScore}(e_j))}{|E_p|}$
    end for
    $E_p \leftarrow$ select_disambiguited_entity_set(G)
    $document_{annotations} = adding(E_p)$
end for

3.3. Indexing

Our approach needs an appropriate indexing method that takes all needed information and statistics given by our entity linking method (EL4DT) into consideration. For this reason, our inverted index method performs the indexing task, which prepares all the information necessary for the benefit of our retrieval and ranking method. Figure (3.3) illustrates the structure of the index performed by our inverted index method.

$$\mathrm{e}_1 +> docNo \rightarrow EOccNbD/pargNo/NbEp/isStrong/NbSEp/NbRE..$$
$$\mathrm{e}_2 +> docNo \rightarrow EOccNbD/pargNo/NbEp/isStrong/NbSEp/NbRE..$$
$$\vdots$$
$$\mathrm{e}_i +> docNo \rightarrow EOccNbD/pargNo/NbEp/isStrong/NbSEp/NbRE..$$
$$\vdots$$
$$\mathrm{e}_n +> docNo \rightarrow EOccNbD/pargNo/NbEp/isStrong/NbSEp/NbRE..$$

Figure 3.3. *Structure of the index*

where each line corresponds to an entity $\boldsymbol{e_i}$ with all documents in which it occurs and all other details. Here, **docNo** represents the document's identifier, **EOccNbD** represents the entity occurrence number in the document, **pargNo** is the paragraph's identifier, **NbEp** represents the number of entities in the paragraph, and **isStrong** takes values (0 or 1), which stands for a non-strong and strong entity, respectively. **NbSEp** represents the number of strong entities in the paragraph, and **NbRE** is the number of semantically related entities identified in entity our linking method.

3.4. Retrieving and Ranking

This session introduces the retrieving and ranking method we designed for our approach as one of its key elements. The retrieval process of the method is an end-to-end process that retrieves all relevant documents for a given query. Once all relevant documents are retrieved, the ranking function of the method ranks documents based on its scoring system, as explained in the following session.

3.4.1. Document Scoring

Computing document scores for a given query (completely annotated query) according to our approach needs an appropriate and relevant solution. To achieve this goal, we design and develop the following ranking method, which is mainly based on the following Equation (3.7). The formula allows to compute the document relevance score, by summing the relevance score of each paragraph in the document against query entities.

$$S(q,d) = \sum_{p\in d} \frac{\sum_{e\in E_q}[nb\_rE(e)\times nbT(e)]\times |E_{qp}|^2 \times e^{[|SE_{qp}|]}}{|E_p| + e^{[|SE_p| - |SE_{qp}|]}} \tag{3.7}$$

where $\boldsymbol{nb\_rE(e)}$ gives the number of a related entity (from the index); $\boldsymbol{nbT(e)}$ gives the number of terms in entity $\boldsymbol{e}$. $|\boldsymbol{E_{qp}}|$, the number of query entities found in paragraph p. $\boldsymbol{e^{[|SE_{qp}|]}}$, the exponential function here, is used to weight the number of query entities located in paragraph p as strong entities according to the index information.

3.4.2. Title Weighting

The document title weight is computed according to Equation (3.8):

$$TW(q,E_t) = |E_{qt}| \times S(q,d) \times w, \tag{3.8}$$

where $|\boldsymbol{E_{qt}}|$ is the number of query entities present in the document title, and $\boldsymbol{w}$ is a parameter used to balance the influence of title weight in the document scoring process. Its value is an arbitrary value established after many tests ($\boldsymbol{w = 0.01}$). Finally, the document title weight is added to the document score.

3.4.3. Algorithm

In this session, we provide the algorithm of our document retrieval and ranking method (Algorithm 2) highlighting all main details. The algorithm represents how to compute the ranking score for retrieved documents corresponding to a given query, where only completely annotated queries are considered.

The inputs are E_p which represents the entities of the given query; subIndexAsRaws represents the loaded lines from the index corresponding to each entity of the query. In line three (3) of the algorithm, getStatisctics() function extracts all statistics and information from the raw lines of the subIndexAsRaws. In line (4) getAllFoundDocsIDs() function retrieves all documents that contain at least one entity of the given query entities, which are the concerned documents. The rest of the algorithm shows how the ranking score is computed for each pair document query. Finally, the algorithm returns $ScoreQ\_docs$, which represents the document ranked list with a ranking score for each retrieved document.

ALGORITHM 2: Retrieval and Ranking Method

1: **Input**: q, E_p, subIndexAsRaws
2: **Output**: $ScoreQ\_docs$
3: entity_index_info ← getStatisctics(subIndexAsRaws)
4: retrievedDocs ← getAllFoundDocsIDs(entity_index_info)
5: **for** $d \in$ retrievedDocs **do**
6: $ScoreQ\_docs(q,d) \leftarrow \sum_{p \in d} \frac{\sum_{e \in E_q}[nb\_rE(e) \times nbT(e)] \times |E_{qp}|^2 \times e^{[|SE_{qp}|]}}{|E_p| + e^{[|SE_p| - |SE_{qp}|]}}$
7: $TW(q, E_t) \leftarrow |E_{qt}| \times ScoreQ\_docs(q,d) \times w$
8: $ScoreQ\_docs(q,d) \leftarrow ScoreQ\_docs(q,d) + TW(q, E_t)$
9: **end for**

3.5. Summary

This chapter presents our approach (PESS4IR). It begins by briefly providing the text preprocessing method. Then it provides our entity linking method (EL4DT), by presenting its details. Then it provides the indexing method. The last section presents the details of the retrieval and ranking method, which is the core of PESS4IR.

4. EXPERIMENTS

In this Chapter, we provide the data collections, evaluation metrics, and some implementation details used for conducting experiments. Then, we present the results and the discussion of the results.

4.1. Data

In our experiments, we use the standard TREC 2004 Robust collection, which was used in TREC 2004 Robust Track. Also, we use the MS MACRO collection (Bajaj, 2016), which is a large-scale dataset focused on machine reading comprehension, question answering, and passage/document ranking. Table 4.1 shows information about our use of these collections.

Table 4.1. *Usage of MSMARCO and Robust04 collections.*

Collection	Queries (title only)	#docs	Qrels
TREC Disks 4 & 5 minus CR	TREC 2004 Robust Track, topics 301-450 & 601-700	528k	Complete qrels[2]
Ms Marco v1	TREC-DL-2019 and TREC-DL-2020 (200, 200)	3.2M	ir_datasets (Python API)[3]

4.2. Evaluation Metrics

The evaluation metrics are computed based on the relevance judgments corresponding to each collection. There are two categories of judgments binary and non-binary relevance judgments:

- Metric for binary relevance judgments: the Mean Average Precision (MAP) is used, and it is computed by taking average of AP over all queries.
- Metric for non-binary relevance judgments, the normalized Discounted Cumulative Gain (nDCG) is used (Järvelin, 2002). The nDCG, as it is named, is the normalization of DCG metric, which is computed by:

$$DCG(L) = \sum_{i=1}^{i=n} \frac{r_i}{log_2(i+1)}, \tag{4.1}$$

where r_i is judgments value, and it varies between 0 to 3 (0 for non-relevant, 1 and 2 somehow relevant, 3 for relevant).

$$nDCG(L) = \frac{DCG(L)}{IDCG(L)}, \tag{4.2}$$

[2] https://trec.nist.gov/data/robust/qrels.robust2004.txt
[3] https://ir-datasets.com/msmarco-document.html

where IDCG is the greater DCG score for a particular query among the list of queries. The result of nDCG metric varies from (0 to 1). The more is close to 1 signifies that the ranking system is performing better.

In our experiments, we use the three standard evaluation metrics. NDCG@20, the official TREC Web Track ad-hoc task evaluation metric, is the first one to evaluate the results. The second metric is the mean average precision (MAP) of the top-ranked 1000 documents. The third metric is P@20 which provides the precision of the top 20 retrieved documents. Moreover, regarding the importance of the 5-top ranked document, NDCG@5 is also used as an evaluation metric for ad-hoc document retrieval tasks. It should be noted that the nDCG@5 evaluation metric is used to compare performances in many different ranking models and ad-hoc document retrieval tasks such as (Wu Z. M., 2020), (Wu C. Z., 2021) and (Yang, 2021).

4.3. Query Annotation

Our approach is purely based on entity representation for documents and queries. Assuming that we have the best-designed purely entity-based retrieval system, including the best representation of the document and the best ranking method, if the given query is not completely annotated, the system will not work well because of the ignored term(s) from that query (not-annotated term(s)). Therefore, query annotation is the critical factor in a purely entity-based retrieval system, which is why we consider only completely annotated queries in our experiments. We test our retrieval approach by using two arbitrary entity linking methods for query annotation, including DBpedia Spotlight (Mendes P. N.-S., 2011) and REL (van H. M., 2020). Moreover, the two Python APIs, provided in Table 4.2, are the used implementations corresponding to each of these two entity linking methods. Table (4.2) presents the number of completely annotated queries by each entity linking method.

Table 4.2. *Completely annotated queries by DBpedia Spotlight and REL for Robust collection queries.*

Entity Linking Method	**#completely annotated queries**	**% of completely annotated queries**	**Usage**
DBpedia Spotlight	180	72%	Spotlight Python Library (v0.7) [4]
REL	17	6.8%	Python APl [5]

[4] https://pypi.org/project/spotlight/
[5] https://github.com/informagi/REL

Moreover, to check whether a query was completely annotated, we compare the found mentions with the original query text. Also, the stopwords are not considered if they are not in mentions. Furthermore, the relevance of query annotations could be checked by the average score of query entities' annotation scores.

Furthermore, Table (4.2) contains only the queries completely annotated by both DBpedia Spotlight and REL entity linkers. Hence, the process is applied to all Robust04 queries (250 queries). Also, we note that no changes were made to REL's results or DBpedia Spotlight query annotations. Later in our tests, we classify queries according to the average scores of their annotation scores to show the corresponding performances and to understand how to effectively leverage a purely entity-based approach.

4.4. Results of Experiments on Robust04

In this section, we explain our results by presenting their evaluations. However, before explaining the results, it is crucial to clarify some information about annotation scores achieved by both DBpedia Spotlight and REL entity linkers, where both methods offer scores between 0 and 1). However, the scoring systems and the meaning of the scores are different.

Like the probability logic, the general interpretation is the same for both methods, which states that the closer the annotation score is to 1, the more accurate the annotation is. And vice versa when the annotation score is closer to zero.

We classified the completely annotated queries into four classes according to their annotation average scores regarding each entity linking method. The reason for this classification is to observe the performance of the PESS4IR method while the annotation score increases. Moreover, four classes (four is an arbitrary number) are suitable for the results' readability. So, the four average score classes are (min = 0.65, 0.85, 0.95, 1.00) and (min = 0.50, 0.65, 0.70, 0.75) for DBpedia Spotlight and REL, respectively. Moreover, for each of these average scores, the corresponding query numbers are (154,132,109, and 3) and (12, 9, 4, and 2), respectively, for each method. We note that DBpedia Spotlight tends to assign higher scores than the REL tool; thus, we assigned the average score classes’ values of DBpedia Spotlight higher than those of REL entity linker. Finally, for both entity linkers, the completely annotated queries are separately classified into these different classes. Figure (4.1) shows the performance of both methods (our approach and Galago search engine), according to each class of queries, where for PESS4IR the query subsets are annotated by DBpedia Spotlight entity linker. It is important to note that Galago implements by default the Dirichlet smoothing function, which is a standard method of interpolating between the collection model and the document model.

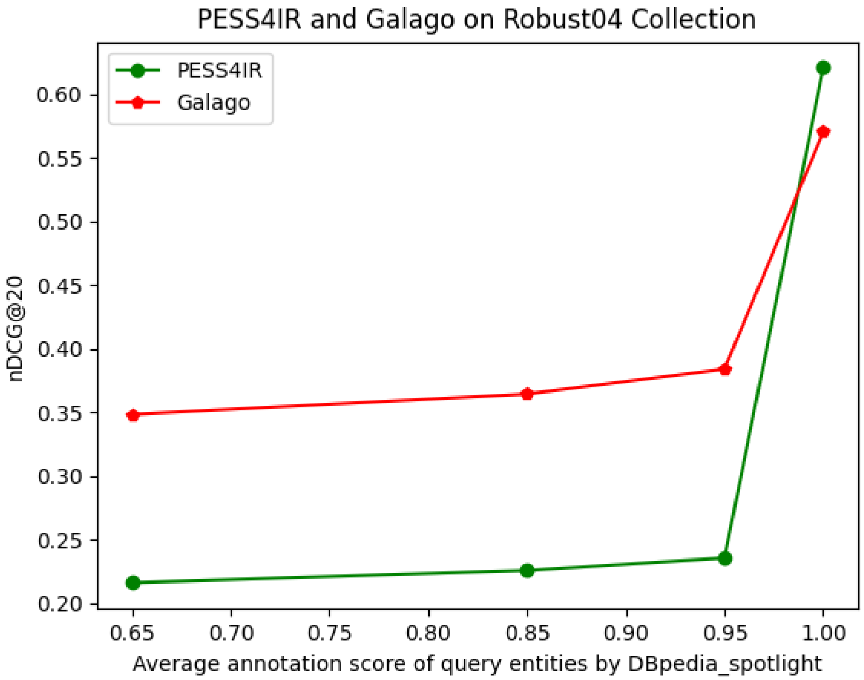

Figure 4.1. *nDCG@20 retrieval scores for the groups of queries annotated by DBpedia spotlight.*

Thus, Galago outperforms our approach for the first three classes of queries. However, for the last query class, PESS4IR beats Galago, where the average annotation score of that class is equal to (1.0). The corresponding nDCG@20 scores are provided in Table (4.3), with more details.

Table 4.3. *nDCG@20 scores for queries' classes given by DBpedia spotlight.*

Method	**nDCG@20**			
	AVGs >= Min (154 queries)	AVGs >=0.85 (132 queries)	AVGs >=0.95 (109 queries)	AVGs =1.0 (3 queries)
Galago (Dirichlet)	0.3498	0.3643	0.3839	0.5702
PESS4IR	0.2160	0.2257	0.2355	**0.6207**

In the same way, Figure (4.2) shows the results for each query class with the corresponding nDCG@20 scores achieved by each method (Galago and PESS4IR), where, in this case, the queries are annotated by the REL entity linking method. Moreover, it can be seen in Figure (4.2) that the PESS4IR's performance continuously increases along with the rise of the query entities' average annotation scores, unlike Galago's performance.

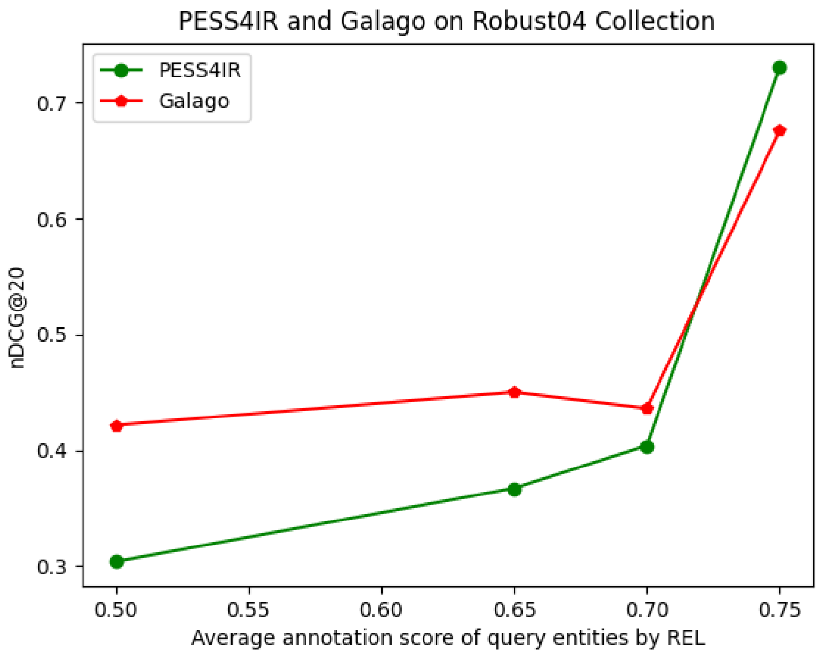

Figure 4.2. *nDCG@20 retrieval scores for the groups of queries annotated by REL.*

Contrary to the three first classes of query entities' average annotation score, in the last queries class, which corresponds to average annotation scores larger than or equal to (0.75), PESS4IR outperforms Galago, and the corresponding nDCG@20 scores are listed in Table (4.4).

Table 4.4. *nDCG@20 scores for queries' classes given by REL.*

Method	nDCG@20			
	AVGs >= Min (12 queries)	AVGs >=0.65 (9 queries)	AVGs >=0.7 (4 queries)	AVGs >=0.75 (2 queries)
Galago (Dirichlet)	0.4216	0.4500	0.4360	0.6759
PESS4IR	0.3036	0.3670	0.4038	**0.7306**

From the perspective of a retrieval system based on multi-method, which leverages different approaches (retrieval method) for better document information retrieval, PESS4IR could be leveraged for the well-represented queries (queries with high annotation scores, such as the class of highest average annotation score, in (Tables 4.3 and 4.4)), and other methods for the rest of queries. In fact, due to its autonomy, the PESS4IR approach could be used alongside any other document retrieval method. Table 4.5 illustrates the added value of PESS4IR when it is used alongside Galago. Moreover, for Galago, the added value is expressed by all used metrics. Furthermore, PESS4IR provides an added value for any state-of-the-art method on the TREC 2004 Robust collection and its query set. When one uses PESS4IR for the highest represented queries set (AVG_score >= 0.75, annotated by REL), and any other state-of-the-art method for the rest of the queries. In this case, an added value is achieved by PESS4IR unless that SOTA method reaches the maximum nDCG@5 score for those highly annotated queries. In the discussion session (see session 4.6.1), we give more details about the added value achieved by PESS4IR.

Table 4.5. *PESS4IR's added value upon Galago.*

Method	**nDCG@5**	**nDCG@20**	**MAP**	**P@20**
Galago (Dirichlet)	0.3729	0.3300	0.1534	0.2795
Galago+PESS4IR	**0.3758**	**0.3311**	**0.1540**	**0.2803**

4.4.1. PESS4IR with LongP (Longformer) Model

We compare PESS4IR against LongP (Longformer) model and combine them to have better performance for ad hoc document retrieval task. Moreover, in this experiment, PESS4IR is used for the highest represented queries set (AVG_score >= 0.75, annotated by REL), and LongP (Longformer) is used for the rest of the queries. The added value achieved by PESS4IR appears when it is used alongside with LongP model. The results are illustrated in the following table (Table 4.6).

Table 4.6. *PESS4IR and Long (Longformer) on Robust Collection.*

Method	**nDCG@5**	**MAP**	**P@5**
Long (Longformer)	0.6542	0.3505	0.6723
Long (Longformer)+PESS4IR	**0.6551**	0.3492	**0.6731**

In Table 4.6, the evaluation is given by nDCG@5 metric, where it shows better ranking performance, which is due to the outperforming of PESS4IR upon Long

(Longformer) model for the highest annotated queries. In session (4.6.1), we provide the details of the added value, achieved by PESS4IR.

4.5. Results of Experiment on MSMARCO

We would test PESS4IR by queries annotated by both REL and DBpedia Spotlight entity linking methods, however, we test it only by the DBpedia Spotlight tool. The reason for not testing PESS4IR with queries annotated by the REL tool is that among the completely annotated queries, of both query sets of TREC-DL-2019 and TREC-DL-2020, there is no query whose annotation avg score is greater than 0.75 (for TREC-DL-2020 query set, see appendix A.3); and for the query set of TREC-DL-2019 there is only one query, but it has a scoring issue (see appendix A.4). In the Discussion (session 4.6.4), we provide the details related to that scoring issue.

In the following table (Table 4.7), we provide the performance and results of our approach (PESS4IR) and LongP (Longformer) model on MSMARCO collection. Moreover, the nDCG@10 metric is used. It is important to note that the Python API of MSMARCO collection does not provide judgment values (qrels) of some queries for TREC DL 2019 and TREC DL 2020 query subsets. Among them, the highest annotated queries exist (annotated by DBpedia Spotlight tool, whose annotation avg score is equal to 1). And with these highest annotated queries PESS4IR supposed to do its best performance. However, the corresponding judgment values are provided by the MSMARCO collection.

Table 4.7. *nDCG@10 scores for PESS4IR and LongP (Longformer) on MSMARCO.*

Method	nDCG@10					
	TREC DL 2019			TREC DL 2020		
	avg>=Min (24 queries)	avg>=0.95 (16 queries)	avg=1.0 (1 query)	avg>=Min (24 queries)	avg>=0.95 (10 queries)	avg=1.0 (1 query)
LongP (Longformer)	**0.7179**	**0.7464**	None	**0.6850**	**0.6605**	None
PESS4IR	0.3842	0.4110	None	0.2733	0.2790	None

The LongP (Longformer) model outperforms PESS4IR for the first two classes for each set (TREC DL 2019 and TREC DL 2020). And for the last class, where PESS4IR is supposed to achieve its best performance with the queries with the highest annotation, there are no corresponding qrels; this why we put “None”. Consequently, the absence of highly annotated queries (with an annotation avg ≥ 0.75) annotated by the REL tool and the absence of corresponding qrels for the highly annotated queries annotated by the DBpedia Spotlight tool (with an annotation avg = 1.0) are the two reasons why we did not provide the combination of both PESS4IR and Long on the MSMARCO collection.

4.6. Discussion

In the discussion session, we discuss our experiments presenting the strengths and limitations of our approach (PESS4IR). Firstly, we explain and discuss the added value achieved by PESS4IR, when it is tested with queries annotated by REL and DBpedia Spotlight entity linking methods. Finally, we discuss some potential strengths of PESS4IR in related tasks such as query performance prediction (QPP). We note that QPP is not the primary concern of this thesis.

4.6.1. PESS4IR Tested by REL:

Since a purely entity-based method is appropriate for only completely annotated queries, the results are partial, where only completely annotated queries are considered. Among them, the ones with higher scores are doing better than the rest. Thus, the purely entity-based approach is recommended for highly represented queries (whose entities have high annotation scores). Furthermore, this session shows how our approach achieves the maximum nDCG@5 score. The following experiment shows how our approach offers added value. In the experiment, we use the REL entity linking method for query annotation (see appendix A.2). Table 4.8 contains highly represented queries whose entities have an average annotation score greater than or equal to 0.75. The table also contains the query text which is the original text (query title). It shows for each query, in REL annotation column, the detected mentions with the corresponding entities and their annotation scores.

Table 4.8. *REL annotations whose average scores are higher than or equal to 0.75.*

qID	Query text	REL annotations (“mention” → entity → score)	Annotation avg scores
365	El Nino	“El Nino”→El_Niño→0.76	0.76
423	Milosevic, Mirjana Markovic	“Milosevic”→Slobodan_Milošević→0.99 “Mirjana Markovic”→Mirjana_Marković→0.98	0.985

In addition, to analyze the performance achieved by our purely entity-based approach, we present the results of PESS4IR together with the results of LongP

(Longformer) and the Galago (Dirichlet) models. Moreover, Figure (4.3) comparatively shows the results of our approach, against the two models. In the experiment, illustrated in Figure (4.3) and Table (4.9), the nDCG@k scores are for the 5-top ranked documents and the results of the LongP (Longformer) model are provided.

With these results, our approach achieves the maximum nDCG@5 score of 1.000 for the highest represented queries (annotated by the REL entity linking method). This score is an added value for any document retrieval method which does not reach that score. Moreover, in this experiment, the achieved score is the maximum nDCG@5 score, corresponding to only two queries. This low number of queries represents a limitation of our approach. But this limitation is caused by comparison needs, where we get that maximum nDCG@5 score after selecting the highest represented queries (with the highest annotation average score). Thus, this is how we compared our approach to any other document retrieval approach.

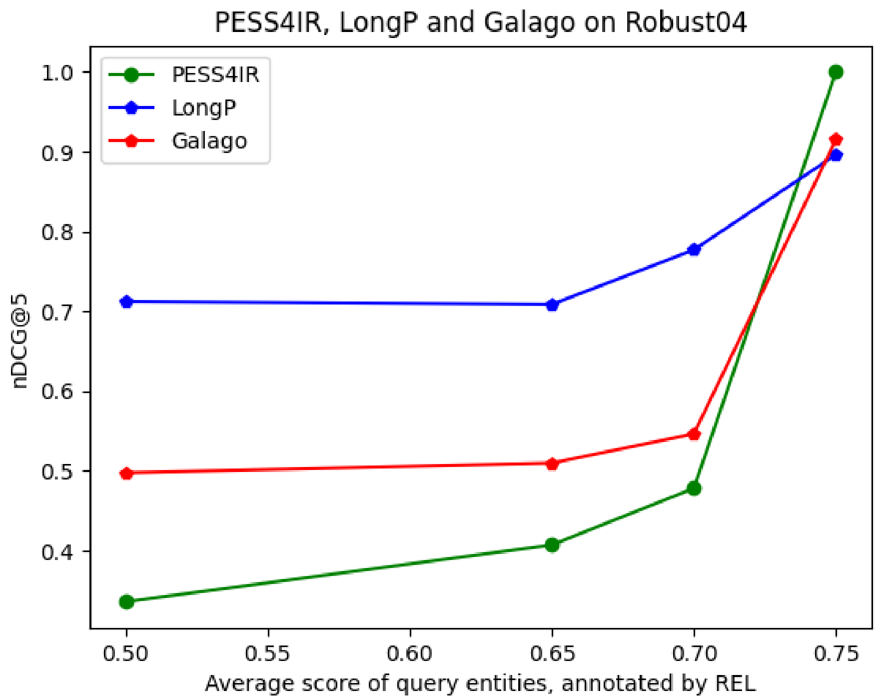

Figure 4.3. *nDCG@5 retrieval scores for the groups of queries annotated by REL.*

Table 4.9 shows how LongP (Longformer) model outperforms PESS4IR and Galago (Dirichlet method), with big difference, for the first three classes of queries. However, queries of the last class (queries whose entities have an average annotation score greater than or equal to 0.75), LongP (Longformer) model is outperformed by both

Galago (Dirichlet method) and PESS4IR, with a domination of our approach by performing the maximum nDCG@5 score.

Table 4.9. *nDCG@5 scores for queries' classes given by REL.*

Method	nDCG@5			
	AVGs >= Min (12 queries)	AVGs >=0.65 (9 queries)	AVGs >=0.7 (4 queries)	AVGs >=0.75 (2 queries)
Galago (Dirichlet)	0.4976	0.5097	0.5463	0.9152
LongP (Longformer)	**0.7122**	**0.7085**	**0.7769**	0.8962
PESS4IR	0.3336	0.4071	0.4781	**1.0000**

4.6.2. PESS4IR Tested by DBpedia Spotlight

As we mentioned before, the DBpedia Spotlight entity linker tends to assign higher scores, for annotations, than the REL entity linking method, which is the reason why the highest represented queries group has a high value. Table 4.9 contains highly represented queries whose entities have an average annotation score equal to 1.0. These queries are the highest represented queries according to the DBpedia Spotlight entity linker. Table 4.10 shows for each query, whose annotation average score equal to 1, the detected mentions with the corresponding entities and their annotation scores.

Table 4.10. *DBpedia Spotlight annotations whose average scores equal to 1.0.*

qID	Query text	DBpedia Spotlight annotations ("mention" → entity → score)	Annotation avg scores
396	sick building syndrome	"sick building syndrome"→ Sick_building_syndrome→1.0	1.0
400	Amazon rain forest	"Amazon rain forest"→ Amazon_rainforest→1.0	1.0
429	Legionnaires' disease	"Legionnaires' disease"→ Legionnaires'_disease→1.0	1.0

Moreover, the performance of the test performed with the DBpedia Spotlight entity linking method corresponding to nDCG@5 scores is presented in Table (4.11). As we can see in the table of achieved performance (Table 4.11), PESS4IR with the queries annotated by DBpedia Spotlight, achieves an added value for Galago with the higher represented queries; however, it is outperformed by LongP (Longformer) model for all queries classes. Thus, when PESS4IR is tested on the Robust04 collection, by queries annotated by the DBpedia Spotlight tool, it does not outperform LongP (Longformer) model.

Table 4.11. *nDCG@5 scores for queries' classes given by DBpedia spotlight*

Method	nDCG@5			
	AVGs >= Min (154 queries)	AVGs >=0.85 (132 queries)	AVGs >=0.95 (109 queries)	AVGs >=1.0 (3 queries)
Galago (Dirichlet)	0.4001	0.4138	0.4392	0.6667
LongP (Longformer)	**0.6737**	**0.6896**	**0.6946**	**0.8284**
PESS4IR	0.2599	0.2739	0.2864	0.7380

4.6.3. Query Performance

The consistency in the relationship between the average annotation scores of queries and their ranking performance is apparent. One can observe clearly from the results listed in Tables 4.7 and 4.9 that, with PESS4IR, the higher the average annotation score of the query gets, the ranking performance becomes better, as well. Based on this clear evidence, with PESS4IR, the query performance is predicted simply by the average annotation score of the given query. In the literature, query performance prediction is an information retrieval task, which has mainly two types of approaches such as pre-retrieval (Zendel, 2021) and post-retrieval (Datta, 2022) approach. Although QPP is not the primary concern of this paper, we think that it is worth to note that, without developing a QPP method, PESS4IR predicts the query performance simply by the query annotation scores.

Based on our experiments, for each entity linking method, we provide below the average annotation score, for which PESS4IR is supposed to do its best.

- REL: For the REL entity linking method, the best performance is achieved with an average annotation score of 0.75. With TREC 2004 Robust collection queries set, the number of completely annotated queries is 17 (Table 4.2), where two have an average score higher than or equal to 0.75. In other words, among the queries completely annotated by the REL entity linker, 11.7% achieved the best possible performance with PESS4IR (see Tables 4.4 and 4.7).
- DBpedia spotlight: For the DBpedia spotlight entity linking method, the best performance is achieved with an average annotation score of 1.0. With

TREC 2004 Robust collection queries set, the number of completely annotated queries is 180 (Table 4), where three have an average score higher than or equal to 1.0. In other words, among the queries completely annotated by the DBpedia spotlight entity linker, 1.6% of them achieved the best possible performance with PESS4IR (see Tables 4.3 and 4.9).

4.6.4. Query Annotation Weakness

The weakness of the annotation of a given query could be represented by using DBpedia spotlight, REL, and TagMe tools, for the query sets such as TREC 2019 and TREC 2020 of MSMACO collection. In the following table (Table 4.12), we show an example of the weakness of a purely based approach, which could be caused by query entity linking methods.

Table 4.12. *Example of query annotation weakness*

qID	Query Text	REL annotations (entity → score)	DBpedia spotlight annotations (entity → score)	TagMe annotations (entity → score)
1037798	who is robert gray	Robert_Gray_(poet) → 0.94	2015_Mississippi_gubernatorial_election→ 0.69	Robert_Gray_(sea_captain) → 0.3

We note that REL gives its annotations of this query as the highest annotated query with an annotation score of 0.94. Such annotation would surely negatively affect any purely entity-based approach. Generally, entity linking methods use a disambiguation process to select an entity among many candidate entities, as we explain in session (3.2.4). In this case, for the given query "*who is robert gray*", among many entities (known people, with the same name) REL method selects "*Robert_Gray_(poet)*" to be the entity. However, the issue is the computed annotation score, which means that the selected entity is sure. In other words, the 0.94 score means that there is no way it could be another "*robert gray*". Thus, such a case would negatively affect our approach and let it have a weak performance when it is supposed to reach its best performance. In the other hand, with TagMe tool the selected entity was "*Robert_Gray_(sea_captain)*", with 0.3 as the annotation score, which is a logical annotation score. We note that PESS4IR gives better score for the annotation of TagMe in this specific case.

4.7. Summary

This chapter presents the performed experiments of our approach PESS4IR. It begins by briefly providing information about the used collections in the Data session. Then it provides the evaluation metrics. Then it provides information about query annotation and the used entity linking tools to achieve it. And then provide the result on Robust and MSMARCO collections. The last section presents the discussions of the results and some important details related to the results.

5. CONCLUSIONS

This chapter provides a summary of our work. We begin by summarizing how a purely entity based could be leveraged for better document retrieval and highlight our primary contributions. Then we conclude by proposing several potential research directions for further research.

5.1. Thesis Summary

We introduce a purely entity-based semantic search approach for ad-hoc information retrieval (PESS4IR) as a novel solution. The main goal of this thesis is to analyze the impact of purely entity-based semantic search on the effectiveness of ad hoc document retrieval by giving clear answers about when such an approach does its best and when it does not, showing its strengths and weaknesses. Our proposed approach represents queries and documents only by entity-based representation. It includes mainly its own entity linking method appropriate for document text (EL4DT), an inverted indexing method, and a method for document retrieving and ranking designed to leverage all strengths of the approach. To evaluate the approach, we used TREC 2004 Robust and MSMARCO collections, and linking query by two different entity linking methods, DBpedia Spotlight and REL. Since our approach uses a purely entity-based representation for queries and documents, only completely annotated queries are considered. Galago (Dirichlet) and LongP (Longformer) models are used to compare the performance on the corresponding group of queries from the two collections, and to show how PESS4IR could be compared to any other retrieval method.

In the experiments, we used the average annotation score of each query's entities as only completely annotated queries are considered. The results indicate that as the average annotation score increases, the ranking score gets higher, as well. Indeed, with the highest-scored queries annotated by DBpedia Spotlight and REL, our approach outperforms the Galago method based on the nDCG@20 evaluation metric. Thus, our approach offers an added value when used with the Galago (Dirichlet method) or LongP (Longformer) model. Moreover, for the queries with the highest annotation average score (avg_score ≥ 0.75) among the queries annotated by the REL entity linking method, our approach achieved the maximum nDCG@5 score (1.000), which would be an added value for any ad-hoc document retrieval method that does not reach the same score for the same queries. LongP (Longformer) model is the confirmation of this added value reached by PESS4IR, where it is among the current state-of-the-art models.

5.2. Future Works

Generally, query entity linking field steel has the necessity to be improved, especially for purely entity-based approaches which need sure annotations. So, for further research, how to increase the quality and number of completely annotated queries can be investigated. It would also be interesting to investigate how to do automatic query reformulation and query recommendation tasks based on knowledge bases, with purely entity representation as output of these techniques.

REFERENCES

Bagheri, E. E.-O. (2018). Neural word and entity embeddings for ad hoc retrieval. *Information Processing & Management*, 54(4), 657-673.

Bajaj, P. C. (2016). Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268.*

Balog, K. (2018). *Entity-oriented search.* Springer Nature, (p. 351).

Beltagy, I. P. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Bollacker, K. E. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. *In Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, (pp. 1247-1250).

Bordes, A. U.-D. (2013). Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Boytsov, L. L. (2022). Understanding performance of long-document ranking models through comprehensive evaluation and leaderboarding. *arXiv preprint arXiv:2207.01262*.

Boytsov, L. L. (2022). Understanding performance of long-document ranking models through comprehensive evaluation and leaderboarding. *arXiv preprint arXiv:2207.01262*.

Brandão, W. C. (2014). Learning to expand queries using entities. *Journal of the Association for Information Science and Technology*, 65(9), 1870-1883.

Chen, L. L. (2018). Short text entity linking with fine-grained topics. *In Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, (pp. 457-466).

Croft, W. B. (2010). Search engines: Information retrieval in practice. (Vol. 520, pp. 131-141).

Dai, Z., & & Callan, J. (2019). Deeper text understanding for IR with contextual neural language modeling. *In Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval (pp. 985-988).*

Dalton, J. D. (2014). Entity query feature expansion using knowledge base links. *In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, (pp. 365-374).

Datta, S. G. (2022). A Relative Information Gain-based Query Performance Prediction Framework with Generated Query Variants. *ACM Transactions on Information Systems (TOIS).*

Dehghani, M. Z. (2017). Neural ranking models with weak supervision. *In Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, (pp. 65-74).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dietz, L. K. (2018). Utilizing knowledge graphs for text-centric information retrieval. *In The 41st international ACM SIGIR conference on research & development in information retrieval*, (pp. 1387-1390).

Ferragina, P. &. (2010). Tagme: on-the-fly annotation of short text fragments (by Wikipedia entities). *In Proceedings of the 19th ACM international conference on information and knowledge management*, (pp. 1625-1628).

Ferragina, P., & Scaiella, U. (2010). Tagme: on-the-fly annotation of short text fragments (by Wikipedia entities). *In Proceedings of the 19th ACM international conference on information and knowledge management*, (pp. 1625-1628).

Gabrilovich, E. R. (2013). Facc1: Freebase annotation of clueweb corpora, version 1 (release date 2013-06-26, format version 1, correction level 0). *Note: http://lemurproject. org/clueweb09/FACC1/Cited by*, 5, 140.

Gao, L., & Callan, J. (2022). Long Document Re-ranking with Modular Re-ranker. *In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2371-2376).*

Gerritse, E. J. (2022). Entity-aware Transformers for Entity Search. *arXiv preprint arXiv:2205.00820*.

Guo, J. Y. (2020). A deep look into neural ranking models for information retrieval. *Information Processing & Management*, 57(6), 102067.

Järvelin, K. &. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS), 20(4), 422-446*.

Kwon, S. O. (2021). Word sense disambiguation based on context selection using knowledge-based word similarity. *Information Processing & Management*, 58(4), 102551.

Lashkari, F. B. (2019). Neural embedding-based indices for semantic search. *Information Processing & Management*, 56(3), 733-755.

Lehmann, J. I. (2015). Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2), 167-195.

Li, C., Yates, A., MacAvaney, S., He, B., & Sun, Y. (2020). Parade: Passage representation aggregation for document reranking. *ACM Transactions on Information Systems*.

Lin, Y. L. (2015). Learning entity and relation embeddings for knowledge graph completion. . *In Twenty-ninth AAAI conference on artificial intelligence*.

Liu, Z. X. (2018). Entity-duet neural ranking: Understanding the role of knowledge graph semantics in neural information retrieval. *arXiv preprint arXiv:1805.07591*.

Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011). DBpedia spotlight: shedding light on the web of documents. *In Proceedings of the 7th international conference on semantic systems*, (pp. 1-8).

Mendes, P. N.-S. (2011). DBpedia spotlight: shedding light on the web of documents. *In Proceedings of the 7th international conference on semantic systems*, (pp. 1-8).

Pappu, A. B. (2017). Lightweight multilingual entity extraction and linking. *In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, (pp. 365-374).

Piccinno, F. &. (2014). From TagME to WAT: a new entity annotator. *In Proceedings of the first international workshop on Entity recognition & disambiguation*, (pp. 55-62).

Reinanda, R. M. (2020). Knowledge graphs: An information retrieval perspective. *Foundations and Trends® in Information Retrieval*, 14(4), 289-444.

Robertson, S. &. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4), 333-389.

Schuhmacher, M. D. (2015). Ranking entities for web queries through text and knowledge. *In Proceedings of the 24th ACM international on conference on information and knowledge management*, (pp. 1461-1470).

Sidi, M., & Gunal, S. (2023). A Purely Entity-Based Semantic Search Approach for Document Retrieval. *Applied Sciences*, 13(18):10285. https://doi.org/10.3390/app131810285.

van, H. M. (2020). Rel: An entity linker standing on the shoulders of giants. *In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 2197-2200).

van, H. M., Hasibi, F., Dercksen, K., Balog, K., & de Vries, A. P. (2020). Rel: An entity linker standing on the shoulders of giants. *In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 2197-2200).

Wang, X., Macdonald, C., Tonellotto, N., & Ounis, I. (2023). ColBERT-PRF: Semantic pseudo-relevance feedback for dense passage and document retrieval. *ACM Transactions on the Web, 17(1), 1-39*.

Wu, C. Z. (2021). Are Neural Ranking Models Robust? . *arXiv preprint arXiv:2108.05018*.

Wu, C. Z. (2022). Are neural ranking models robust? *ACM Transactions on Information Systems, 41(2), 1-36*.

Wu, Z. M. (2020). Leveraging passage-level cumulative gain for document ranking. *In Proceedings of The Web Conference 2020*, (pp. 2421-2431).

Xiong, C. C. (2016). Bag-of-entities representation for ranking. *In Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, (pp. 181-184).

Xiong, C. C. (2017). Word-entity duet representations for document ranking. *In Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval*, (pp. 763-772).

Xiong, C. D. (2017). End-to-end neural ad-hoc ranking with kernel pooling. *In Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval*, (pp. 55-64).

Yang, T. &. (2021). Maximizing marginal fairness for dynamic learning to rank. *In Proceedings of the Web Conference 2021*, (pp. 137-145).

Zamani, H. D.-M. (2018). From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. *In Proceedings of the 27th ACM international conference on information*, (pp. 497-506).

Zendel, O. C. (2021). Is Query Performance Prediction With Multiple Query Variations Harder Than Topic Performance Prediction? *In Proceedings of the 44th*

International ACM SIGIR Conference on Research and Development in Information Retrieval, (pp. 1713-1717).

APPENDIX

In this section, we provide the query annotations used to test our approach. The annotations of query sets are achieved by two entity linking methods DBpedia Spotlight and REL.

A.1. DBpedia Spotlight Annotations for Robust04

In the appendix section, we have a query set of the TREC 2004 Robust collection, annotated by the DBpedia Spotlight entity linker. Here, we have only the completely annotated queries. For each query, the average score is provided.

qID<++> Y<++>AVG_score<++>Annotations
301<++>Y<++>0.9422930034709613<++>International_law->Organized_crime
303<++>Y<++>0.9916617532774215<++>Hubble_Space_Telescope->Xbox_Live
305<++>Y<++>0.8233757355772809<++>Bridge_of_Independent_Lists->Dangerous_(Michael_Jackson_album)->Vehicle
308<++>Y<++>0.999981850897192<++>Dental_implant->Dentistry
309<++>Y<++>0.8381166338949552<++>Rapping->Crime
310<++>Y<++>0.9636187527358652<++>Radio_Waves_(Roger_Waters_song)->Brain->Cancer
311<++>Y<++>0.9999999998394102<++>Industrial_espionage
312<++>Y<++>0.9999998935852566<++>Hydroponics
314<++>Y<++>0.9637588769999175<++>United_States_Marine_Corps->Vegetation
316<++>Y<++>0.9823531973603806<++>Polygamy->Polyandry->Polygyny
321<++>Y<++>0.8304033796129933<++>Woman->Parliament_of_England
322<++>Y<++>0.9761505135024882<++>International_law->Art->Crime
323<++>Y<++>0.9989506398073358<++>Literature->Journalism->Plagiarism
324<++>Y<++>0.843523719434736<++>Argentina->United_Kingdom->International_relations
325<++>Y<++>0.9957677409995997<++>Cult->Lifestyle_(sociology)
327<++>Y<++>0.6741173791837178<++>Modern_architecture->Slavery
329<++>Y<++>0.9026182851898723<++>Mexico->Air_pollution
331<++>Y<++>0.9392907092908471<++>World_Bank->Criticism
332<++>Y<++>0.9928067801874498<++>Income_tax->Tax_evasion
333<++>Y<++>0.9998904550378483<++>Antibiotic->Bacteria->Disease
334<++>Y<++>0.9953981065544416<++>Export->Control_system->Cryptography
336<++>Y<++>0.8170574260324551<++>Race_and_ethnicity_in_the_United_States_Census->Bear->Weather_Underground
337<++>Y<++>0.9999999999997335<++>Viral_hepatitis
338<++>Y<++>0.9999863000468299<++>Risk->Aspirin
340<++>Y<++>0.7146518568004271<++>Land->Mining->Ban_of_Croatia
341<++>Y<++>0.9999999992114041<++>Airport_security
342<++>Y<++>0.6708548569598859<++>Diplomacy->Expulsion_of_the_Acadians
343<++>Y<++>0.9932852359003905<++>Police->Death
346<++>Y<++>0.984505597445497<++>Education->Technical_standard
347<++>Y<++>0.9994111465790465<++>Wildlife->Extinction
348<++>Y<++>0.99999987750514<++>Agoraphobia
349<++>Y<++>0.9992382152114924<++>Metabolism
350<++>Y<++>0.9953751424684443<++>Health->Computer->Airport_terminal
351<++>Y<++>0.9527758884363138<++>Falkland_Islands->Petroleum->Hydrocarbon_exploration
352<++>Y<++>0.8502584285986691<++>United_Kingdom->Channel_Tunnel->Impact_event
353<++>Y<++>0.9723341881170074<++>Antarctica->Exploration
354<++>Y<++>0.9620560629515208<++>Journalist->Risk
356<++>Y<++>0.896155611833978<++>Menopause->Estrogen->United_Kingdom
357<++>Y<++>0.8588779634116539<++>Territorial_waters->Sea_of_Japan_naming_dispute
358<++>Y<++>0.9882173961307686<++>Blood_alcohol_content->Death
360<++>Y<++>0.8809526917328019<++>Drug_liberalization->Employee_benefits

361<++>Y<++>0.995345089861352<++>Clothing->Sweatshop
362<++>Y<++>0.8963027195944302<++>People_smuggling
363<++>Y<++>0.9956160648827447<++>Transport->Tunnel->Disaster
364<++>Y<++>0.9982317779257299<++>Rabies
365<++>Y<++>0.9716041526723712<++>El_Niño
367<++>Y<++>0.9692354504948936<++>Piracy
369<++>Y<++>0.9999999999999822<++>Anorexia_nervosa->Bulimia_nervosa
370<++>Y<++>0.9988901469768043<++>Food->Prohibition_of_drugs
371<++>Y<++>0.9276354193704037<++>Health_insurance->Holism
372<++>Y<++>0.9983551804874915<++>Native_American_gaming->Casino
374<++>Y<++>0.9685768957420315<++>Nobel_Prize->Fields_Medal
375<++>Y<++>0.9999999999838174<++>Hydrogen_fuel
376<++>Y<++>0.8702255291357396<++>International_Court_of_Justice
377<++>Y<++>0.9713341665095577<++>Cigar->Smoking
379<++>Y<++>0.9852618198777502<++>Mainstreaming_(education)
380<++>Y<++>0.9595951291022093<++>Obesity->Therapy
381<++>Y<++>0.9999912147260778<++>Alternative_medicine
382<++>Y<++>0.9818197972672459<++>Hydrogen->Fuel->Car
383<++>Y<++>0.8474594732725742<++>Mental_disorder->Drug
384<++>Y<++>0.6671967503078107<++>Outer_space->Train_station->Moon
385<++>Y<++>0.8686428839939203<++>Hybrid_electric_vehicle->Fuel->Car
387<++>Y<++>0.9988472933852381<++>Radioactive_waste
388<++>Y<++>0.9999914286894456<++>Soil->Human_enhancement
389<++>Y<++>0.664255919865177<++>Law->Technology_transfer
390<++>Y<++>0.9999999999991616<++>Orphan_drug
391<++>Y<++>0.9999284901225709<++>Research_and_development->Prescription_costs
392<++>Y<++>0.9995912495852758<++>Robotics
393<++>Y<++>0.9999999999130935<++>Euthanasia
395<++>Y<++>0.9997553022202351<++>Tourism
396<++>Y<++>1.0<++>Sick_building_syndrome
397<++>Y<++>0.9990813178361907<++>Car->Product_recall
400<++>Y<++>1.0<++>Amazon_rainforest
402<++>Y<++>0.9999999999781828<++>Behavioural_genetics
403<++>Y<++>0.999999813435991<++>Osteoporosis
404<++>Y<++>0.6941772057336428<++>Ireland->Peace->Camp_David_Accords
405<++>Y<++>0.4238116228174884<++>Cosmic_ray->Event-driven_programming
407<++>Y<++>0.9923802512157526<++>Poaching->Wildlife->Fruit_preserves
408<++>Y<++>0.9988554001947672<++>Tropical_cyclone
410<++>Y<++>0.9999999999999503<++>Schengen_Agreement
411<++>Y<++>0.9947331398435401<++>Marine_salvage->Shipwreck->Treasure
412<++>Y<++>0.9999999992114041<++>Airport_security
413<++>Y<++>0.9638309080048731<++>Steel->Record_producer
414<++>Y<++>0.9965999250683589<++>Cuba->Sugar->Export
415<++>Y<++>0.775328268440912<++>Drug->Golden_Triangle_of_Jakarta
416<++>Y<++>0.9089337936090394<++>Three_Gorges->Project
419<++>Y<++>0.9917482813095554<++>Recycling->Car->Tire
420<++>Y<++>0.9955077748807217<++>Carbon_monoxide_poisoning
421<++>Y<++>0.988845290029708<++>Industrial_waste->Waste_management
423<++>Y<++>0.9893092495209957<++>Slobodan_Milošević->Mirjana_Marković
424<++>Y<++>0.9964270526243968<++>Suicide
425<++>Y<++>0.9999999999996945<++>Counterfeit_money
426<++>Y<++>0.8827453155184075<++>Law_enforcement->Dog
427<++>Y<++>0.7088978447187699<++>Ultraviolet->Damages->Human_eye
428<++>Y<++>0.983580647717166<++>Declension->Birth_rate
429<++>Y<++>1.0<++>Legionnaires'_disease
430<++>Y<++>0.736355241590634<++>Africanized_bee->September_11_attacks
431<++>Y<++>0.9939081414531497<++>Robotics->Technology
432<++>Y<++>0.9928793873474029<++>Racial_profiling->Driving->Police
433<++>Y<++>0.9999999990127844<++>Ancient_Greek_philosophy->Stoicism

434<++>Y<++>0.9914165431145454<++>Estonia->Economy
435<++>Y<++>0.9997750088796703<++>Curb_stomp->Population_growth
436<++>Y<++>0.8336677830661147<++>Classification_of_railway_accidents
437<++>Y<++>0.8809029694466801<++>Deregulation->Natural_gas->Electricity
439<++>Y<++>0.9930600215575294<++>Invention->Science_and_technology_in_the_Philippines
440<++>Y<++>0.9986339435196137<++>Child_labour
441<++>Y<++>0.9999999999999893<++>Lyme_disease
443<++>Y<++>0.9957246203674307<++>United_States->Investment->Africa
444<++>Y<++>0.9999999999999964<++>Supercritical_fluid
447<++>Y<++>0.9999999999975735<++>Stirling_engine
450<++>Y<++>0.9937577543728069<++>Hussein_of_Jordan->Peace
601<++>Y<++>0.9971377057112235<++>Turkey->Iraq->Water
602<++>Y<++>0.9984578739512397<++>0.6696008778483643<++>Czech_language->Slovakia->Sovereignty
603<++>Y<++>0.9999626386064216<++>0.9985629347827838<++>Tobacco->Cigarette->Lawsuit
604<++>Y<++>0.9999235578240981<++>Lyme_disease->Arthritis
605<++>Y<++>0.9263050230611971<++>Great_Britain->Health_care
606<++>Y<++>0.7390800894132427<++>Human_leg->Trapping->Ban_of_Croatia
607<++>Y<++>0.9965927163010586<++>Human->Genetic_code
609<++>Y<++>0.9920468274116302<++>Per_capita->Alcoholic_drink
610<++>Y<++>0.6887291050943438<++>Minimum_wage->Adverse_effect->Impact_event
611<++>Y<++>0.9944237923763072<++>Kurds->Germany->Violence
612<++>Y<++>0.863878896730292<++>Tibet->Protest
613<++>Y<++>0.7739763636616234<++>Berlin->Berlin_Wall->Waste_management
614<++>Y<++>0.9101682857931109<++>Flavr_Savr->Tomato
615<++>Y<++>0.9997069460982296<++>Lumber->Export->Asia
616<++>Y<++>0.9976499909670737<++>Volkswagen->Mexico
617<++>Y<++>0.9915648387755583<++>Russia->Cuba->Economy
619<++>Y<++>0.9901288174962835<++>Winnie_Madikizela-Mandela->Scandal
620<++>Y<++>0.9954808229883216<++>France->Nuclear_weapons_testing
622<++>Y<++>0.9999999999172893<++>Price_fixing
623<++>Y<++>0.9885496976198986<++>Toxicity->Chemical_weapon
624<++>Y<++>0.8927872609865086<++>Strategic_Defense_Initiative->Star_Wars
625<++>Y<++>0.9703964319776107<++>Arrest->Bomb->World_Triathlon_Corporation
626<++>Y<++>0.999999238626556<++>Stampede
628<++>Y<++>0.9156726801921176<++>United_States_invasion_of_Panama->Panama
629<++>Y<++>0.8864125697999727<++>Abortion_clinic->Attack_on_Pearl_Harbor
630<++>Y<++>0.9999999999999929<++>Gulf_War_syndrome
632<++>Y<++>0.7594953405841971<++>Southeast_Asia->Tin
633<++>Y<++>0.9999999999956017<++>Devolution_in_the_United_Kingdom
635<++>Y<++>0.9791804337848896<++>Physician->Assisted_suicide->Suicide
638<++>Y<++>0.9999999999920917<++>Miscarriage_of_justice
640<++>Y<++>0.9772947307709348<++>Parental_leave->Policy
641<++>Y<++>0.7974386442056666<++>Exxon_Valdez->Wildlife->Marine_life
642<++>Y<++>0.9293590486123976<++>Tiananmen_Square->Protest
643<++>Y<++>0.9958501365753133<++>Salmon->Dam->Pacific_Northwest
644<++>Y<++>0.8128402445905525<++>Introduced_species->Import
645<++>Y<++>0.9999999999699298<++>Copyright_infringement
648<++>Y<++>0.994918609349214<++>Parental_leave->Law
649<++>Y<++>0.9999999999584972<++>Computer_virus
650<++>Y<++>0.9960382314988634<++>Tax_evasion->Indictment
651<++>Y<++>0.9949112351673097<++>United_States->Ethnic_group->Population
653<++>Y<++>0.8261480970551885<++>ETA_SA->Basque_language->Terrorism
657<++>Y<++>0.8118982582118629<++>School_prayer->Smoking_ban
658<++>Y<++>0.9980005204988003<++>Teenage_pregnancy
659<++>Y<++>0.9574704050707363<++>Cruise_ship->Health->Safety
660<++>Y<++>0.999429831087146<++>Whale_watching->California
665<++>Y<++>0.9999825174785343<++>Poverty->Africa->Sub-Saharan_Africa
668<++>Y<++>0.998088959251928<++>Poverty->Disease
669<++>Y<++>0.9999828526608379<++>Iranian_Revolution

670<++>Y<++>0.9999998591162672<++>Elections_in_the_United_States->Apathy
675<++>Y<++>0.9023200615457991<++>Olympic_Games->Training->Swimming
676<++>Y<++>0.9024509959024143<++>Poppy->Horticulture
678<++>Y<++>0.8176555408184811<++>Joint_custody->Impact_event
679<++>Y<++>0.7772527227567606<++>Chess_opening->Adoption->Phonograph_record
680<++>Y<++>0.8252586633730941<++>Immigration->Spanish_language->School
681<++>Y<++>0.8076328345732521<++>Wind_power->Location
682<++>Y<++>0.8430780796585148<++>Adult->Immigration->English_language
685<++>Y<++>0.7973786182622121<++>Academy_Awards->Win–loss_record_(pitching)->Natural_selection
686<++>Y<++>0.9410682082027008<++>Argentina->Fixed_exchange-rate_system->Dollar
687<++>Y<++>0.9920209145313614<++>Northern_Ireland->Industry
689<++>Y<++>0.9962950350527093<++>Family_planning->Aid
691<++>Y<++>0.9991775251948098<++>Clearcutting->Forest
693<++>Y<++>0.9997175525795037<++>Newspaper->Electronic_media
694<++>Y<++>0.9999999999999929<++>Compost
695<++>Y<++>0.7501223260163279<++>White-collar_crime->Sentence_(linguistics)
696<++>Y<++>0.9652985448255742<++>Safety->Plastic_surgery
697<++>Y<++>0.9999999999999822<++>Air_traffic_controller
698<++>Y<++>0.9999767970588322<++>Literacy->Africa
699<++>Y<++>0.9217820925410557<++>Term_limit
700<++>Y<++>0.975172236248435<++>Fuel_tax->United_States

A.2. REL Annotations for Robust04

In the appendix section, we have a query set of the TREC 2004 Robust collection, annotated by the REL entity linker. We provide only the completely annotated queries. For each query, the average score is provided. The information in the index are: (qID: query ID; Y: Yes, query completely annotated; AVG_score: computed average score; Query_Annotations: annotations of a query).

qID<++>Y<++>AVG_score<++>Query_Annotations
301<++>Y<++>0.51<++>Transnational_organized_crime
302<++>Y<++>0.515<++>Polio->Post-polio_syndrome
308<++>Y<++>0.74<++>Dental_implant
310<++>Y<++>0.605<++>Radio_wave->Brain_tumor
320<++>Y<++>0.72<++>Submarine_communications_cable
326<++>Y<++>0.59<++>MV_Princess_of_the_Stars
327<++>Y<++>0.56<++>Slavery_in_the_21st_century
341<++>Y<++>0.6<++>Airport_security
348<++>Y<++>0.65<++>Agoraphobia
365<++>Y<++>0.76<++>El_Niño
376<++>Y<++>0.74<++>The_Hague
381<++>Y<++>0.55<++>Alternative_medicine
416<++>Y<++>0.65<++>Three_Gorges_Dam
423<++>Y<++>0.985<++>Slobodan_Milošević->Mirjana_Marković
630<++>Y<++>0.63<++>Gulf_War_syndrome
669<++>Y<++>0.67<++>Iranian_Revolution
677<++>Y<++>0.69<++>Leaning_Tower_of_Pisa

A.3 REL Annotations for TREC DL 2019

In the appendix section, we have the query set of the TREC DL 2019 (MSMARCO collection), annotated by the REL entity linker. We provide only the completely annotated queries. For each query, the average score is provided. The information in the index are: (qID: query ID; Y: Yes, query completely annotated; AVG_score: computed average score; Query_Annotations: annotations of a query).

qID<++>Y<++>AVG_score<++>Query_Annotations
835929<++>Y<++>0.62<++>United_States_presidential_nominating_convention
1037798<++>Y<++>0.94<++>Robert_Gray_(sea_captain)
1115392<++>Y<++>0.29<++>Phillips_Exeter_Academy_Library

A.4 REL Annotations for TREC DL 2020

In the appendix section, we have the query set of the TREC DL 2020 (MSMARCO collection), annotated by the REL entity linker. We provide only the completely annotated queries. For each query, the average score is provided. The information in the index are: (qID: query ID; Y: Yes, query completely annotated; AVG_score: computed average score; Query_Annotations: annotations of a query).

qID<++>Y<++>AVG_score<++>Query_Annotations
985594<++>Y<++>0.54<++>Cambodia
999466<++>Y<++>0.57<++>Velbert

A.5 DBpedia Spotlight Annotations for TREC DL 2019

In the appendix section, we have the query set of the TREC DL 2019 (MSMARCO collection), annotated by the DBpedia Spotlight entity linker. We provide only the completely annotated queries. For each query, the average score is provided. The information in the index are: (qID: query ID; Y: Yes, query completely annotated; AVG_score: computed average score; Query_Annotations: annotations of a query).

qID<++>Y<++>AVG_score<++>Query_Annotations
1127622<++>Y<++>0.8484174385279352<++>Semantics->Heat_capacity
190044<++>Y<++>0.8865360168634105<++>Food->Detoxification->Liver->Nature
264403<++>Y<++>0.7427101323971266<++>Long_jump->Data_recovery->Rhytidectomy->Neck->Elevator
421756<++>Y<++>0.9887430913683066<++>Pro_rata->Newspaper
1111546<++>Y<++>0.8968400528320386<++>Mediumship->Artisan
156493<++>Y<++>0.7635869346703801<++>Goldfish->Evolution
1124145<++>Y<++>0.8279115042507935<++>Truncation->Semantics
1110199<++>Y<++>0.9999999991887911<++>Wi-Fi->Bluetooth
835929<++>Y<++>0.6801064489366196<++>National_Convention
432930<++>Y<++>0.674476942756101<++>JavaScript->Letter_case->Alphabet->String_instrument
1044797<++>Y<++>1.0<++>Non-communicable_disease
1124464<++>Y<++>0.5242881180978325<++>Quad_scull->Casting
130510<++>Y<++>0.9984735189751052<++>Definition->Declaratory_judgment
1127893<++>Y<++>0.9984366536772885<++>Ben_Foster->Association_football->Net_worth
646207<++>Y<++>0.8550360631796995<++>Production_designer->Fee_tail
573724<++>Y<++>0.997942323584422<++>Social_determinants_of_health_in_poverty->Health
1055865<++>Y<++>0.952787250107581<++>African_Americans->Win–loss_record_(pitching)->Wimbledon_F.C.
494835<++>Y<++>0.99176134505693<++>Sensibility->Definition
1126814<++>Y<++>0.9993302443272604<++>Noct->Temperature
100983<++>Y<++>0.9977403165673293<++>Cost->Cremation
1119092<++>Y<++>0.9999999990881676<++>Multi-band_device
1133167<++>Y<++>0.9940850423375566<++>Weather->Jamaica
324211<++>Y<++>0.930982239901244<++>Money->United_Airlines->Sea_captain->Aircraft_pilot
11096<++>Y<++>0.9849797749940885<++>Honda_Integra->Toothed_belt->Replacement_value
1134787<++>Y<++>0.8745724110755091<++>Subroutine->Malt
527433<++>Y<++>0.9537101464933078<++>Data_type->Dysarthria->Cerebral_palsy
694342<++>Y<++>0.9330494647762133<++>Geological_period->Calculus
1125225<++>Y<++>0.814538265672667<++>Chemical_bond->Strike_price
1136427<++>Y<++>0.7061718630217163<++>SATB->Video_game_developer
719381<++>Y<++>0.6677662534973824<++>Arabic->Balance_wheel
131651<++>Y<++>0.9335919424902749<++>Definition->Harmonic
1037798<++>Y<++>0.6999974850327338<++>2015_Mississippi_gubernatorial_election
915593<++>Y<++>0.9148964938941618<++>Data_type->Food->Cooking->Sous-vide
264014<++>Y<++>0.8141469569212276<++>Vowel_length->Biological_life_cycle->Flea
1121402<++>Y<++>0.989264712335901<++>Contour_plowing->Redox
1117099<++>Y<++>0.9999999904273409<++>Convergent_boundary
744366<++>Y<++>0.9999997784843903<++>Epicureanism
277780<++>Y<++>0.999845912562023<++>Calorie->Tablespoon->Mayonnaise
1114563<++>Y<++>0.9999999999999787<++>FTL_Games

903469<++>Y<++>0.9868563759225631<++>Health->Dieting
1112341<++>Y<++>0.9740228833162581<++>Newspaper->Life->Thai_people
706080<++>Y<++>0.9999999999775682<++>Domain_name
1120868<++>Y<++>0.8666884704281476<++>Color->Louisiana->Technology
523270<++>Y<++>0.9978601407237909<++>Toyota->Plane_(tool)->Plane_(tool)->Texas
133358<++>Y<++>0.8321951248053688<++>Definition->Counterfeit->Money
67262<++>Y<++>0.9596081186595659<++>Farang->Album->Thailand
805321<++>Y<++>0.8853931908810876<++>Area->Rock_music->Psychological_stress->Breakbeat-
>Database_trigger->Earthquake
1129828<++>Y<++>0.960301020029886<++>Weighted_arithmetic_mean->Sound_bite
131843<++>Y<++>0.993713148032662<++>Definition->SIGMET
104861<++>Y<++>0.9951204000467133<++>Cost->Interior_design->Concrete->Flooring
833860<++>Y<++>0.9681002268307477<++>Popular_music->Food->Switzerland
207786<++>Y<++>0.9999370910168783<++>Shark->Warm-blooded
691330<++>Y<++>0.9999992829052942<++>Moderation_(statistics)
1103528<++>Y<++>0.9972950550942021<++>Major_League_(film)
1132213<++>Y<++>0.7489801473531366<++>Length_overall->Professional_wrestling_holds-
>Bow_and_arrow->Yoga
1134138<++>Y<++>0.7215343120469786<++>Honorary_degree->Semantics
138632<++>Y<++>0.9113521779260643<++>Definition->Tangent
1114819<++>Y<++>0.9999946476896949<++>Durable_medical_equipment->Train
747511<++>Y<++>0.9999998038955745<++>Firewalking
183378<++>Y<++>0.9989397404012138<++>Exon->Definition->Biology
1117387<++>Y<++>0.8663803334217364<++>Chevy_Chase->Semantics
479871<++>Y<++>0.9503704570127932<++>President_of_the_United_States->Synonym
541571<++>Y<++>0.9983833679282048<++>Wat->Dopamine
1106007<++>Y<++>0.8808753545444665<++>Definition->Visceral_leishmaniasis
60235<++>Y<++>0.836409024736343<++>Calorie->Egg_as_food->Frying
490595<++>Y<++>0.7290108662022954<++>RSA_Security->Definition->Key_size
564054<++>Y<++>0.9999999966859434<++>Red_blood_cell_distribution_width->Blood_test
1116052<++>Y<++>0.8321774517493923<++>Synonym->Thorax
443396<++>Y<++>0.9814649278583856<++>Lipopolysaccharide->Law->Definition
972007<++>Y<++>0.9622847968581714<++>Chicago_White_Sox->Play_(theatre)->Chicago
1133249<++>Y<++>0.7394092678658755<++>Adenosine_triphosphate->Record_producer
101169<++>Y<++>0.9949249424089939<++>Cost->Jet_fuel
19335<++>Y<++>0.8545708866482175<++>Anthropology->Definition->Natural_environment
789700<++>Y<++>0.9999999009245122<++>Resource-based_relative_value_scale
47923<++>Y<++>0.8507968217623343<++>Axon->Nerve->Synapse->Control_knob->Definition
301524<++>Y<++>0.9719576176244117<++>Zero_of_a_function->Names_of_large_numbers
952774<++>Y<++>0.7970879064723523<++>Evening
766511<++>Y<++>0.7354697185453023<++>Lewis_Machine_and_Tool_Company->Stock
452431<++>Y<++>0.9935533902835246<++>Melanoma->Skin_cancer->Symptom
1109818<++>Y<++>0.773903290136571<++>Experience_point->Exile
1047902<++>Y<++>0.9396894541136506<++>Play_(theatre)->Gideon_Fell->The_Vampire_Diaries
662372<++>Y<++>0.8886998123462867<++>Radio_format->USB_flash_drive->Mackintosh
364142<++>Y<++>0.8255594621305994<++>Wound_healing->Delayed_onset_muscle_soreness
20455<++>Y<++>0.9396229761461882<++>Arabic->Glasses->Definition
1126813<++>Y<++>0.7556818914101636<++>Nuclear_Overhauser_effect->Bone_fracture
240053<++>Y<++>0.7554636687709102<++>Vowel_length->Safety->City_council->Class_action->Goods
1122461<++>Y<++>0.9992610139419709<++>Hydrocarbon->Lipid

```
1116341<++>Y<++>0.8146863386208845<++>Closed_set->Armistice_of_11_November_1918->Mortgage_loan->Definition
1129237<++>Y<++>0.9981516927084026<++>Hydrogen->Liquid->Temperature
423273<++>Y<++>0.9999999989010391<++>School_meal->Tax_deduction
321441<++>Y<++>0.9990492057816107<++>Postage_stamp->Cost
```

A.6 DBpedia Spotlight Annotations for TREC DL 2020

In the appendix section, we have the query set of the TREC DL 2019 (MSMARCO collection), annotated by the DBpedia Spotlight entity linker. We provide only the completely annotated queries. For each query, the average score is provided. The information in the index are: (qID: query ID; Y: Yes, query completely annotated; AVG_score: computed average score; Query_Annotations: annotations of a query).

qID<++>Y<++>AVG_score<++>Query_Annotations
1030303<++>Y<++>0.7340946183870847<++>Shaukat_Aziz->Banu_Hashim
1043135<++>Y<++>0.946317312457761<++>Killed_in_action->Nicholas_II_of_Russia->Russia
1045109<++>Y<++>0.7511704665155204<++>Holding_company->John_Hendley_Barnhart->Common_crane
1051399<++>Y<++>0.9831254656185995<++>Singing->Monk->Theme_music
1064670<++>Y<++>0.9970763927894758<++>Hunting->Pattern->Shotgun
1071750<++>Y<++>0.8892464638623135<++>Pete_Rose->Smoking_ban->Hall->Celebrity
1105860<++>Y<++>0.8774266878935226<++>Amazon_rainforest->Location
1106979<++>Y<++>0.9844715509684185<++>Exponentiation->Pareto_chart->Statistics
1108450<++>Y<++>0.8991756241023721<++>Definition->Definition->Gallows
1108466<++>Y<++>0.9749988943992814<++>Connective_tissue->Composer->Subcutaneous_tissue
1108473<++>Y<++>0.8764035354741885<++>Time_zone->Stone_(unit)->Paul_the_Apostle->Minnesota
1108729<++>Y<++>0.9977600686467922<++>Temperature->Humidity->Charcuterie
1109699<++>Y<++>0.99999999999838<++>Mental_disorder
1109707<++>Y<++>0.9340983506154318<++>Transmission_medium->Radio_wave->Travel
1114166<++>Y<++>0.6642622531448081<++>Call_to_the_bar->Blood->Thin_film
1114286<++>Y<++>0.8685393856480332<++>Meat->Group_(mathematics)
1115210<++>Y<++>0.9995145949464947<++>Chaff->Flare
1116380<++>Y<++>0.9239129473029049<++>Unconformity->Earth_science
1119543<++>Y<++>0.7608774457395511<++>Psychology->Cancer_screening->Train->Egg->Organ_donation
1120588<++>Y<++>0.8671935828811231<++>Tooth_decay->Detection->System
1122138<++>Y<++>0.8878011096897418<++>Symptom->Goat
1122767<++>Y<++>0.8876396654279999<++>Amine->Record_producer->Carnitine
1125755<++>Y<++>0.5846332776541447<++>1994_Individual_Speedway_World_Championship->Definition
1127004<++>Y<++>0.9926518244345025<++>Millisecond->Symptom->Millisecond
1127233<++>Y<++>0.8206546286691387<++>Monk->Semantics
1127540<++>Y<++>0.8369614887321695<++>Semantics->Shebang_(Unix)
1128456<++>Y<++>0.9967168751073104<++>Medicine->Ketorolac->Narcotic
1130705<++>Y<++>0.9987446870472948<++>Passport
1130734<++>Y<++>0.9058236234747112<++>Corn_starch->Giraffe->Thickening_agent
1131069<++>Y<++>0.8074044203528561<++>Son->Robert_Kraft
1132044<++>Y<++>0.9849122526400067<++>Brick->Wall
1132247<++>Y<++>0.8942829158806607<++>Vowel_length->Cooking->Potato_wedges->Oven->Frozen_food
1132842<++>Y<++>0.7258539998537346<++>Vowel_length->Stay_of_execution->Infection->Influenza
1132943<++>Y<++>0.8153913684684001<++>Vowel_length->Cooking->Artichoke
1132950<++>Y<++>0.8255429953411267<++>Vowel_length->Hormone->Headache
1133579<++>Y<++>0.9131369795803623<++>Granulation_tissue->Starting_pitcher
1134094<++>Y<++>0.8285001731543475<++>Interagency_hotshot_crew->Member_of_parliament
1134207<++>Y<++>0.9409175836209229<++>Holiday->Definition
1134680<++>Y<++>0.9766952230811329<++>Jenever->Provinces_of_Turkey->Median->Sales->Price
1134939<++>Y<++>0.9912127141822535<++>Overpass->Definition

```
1135268<++>Y<++>0.9793535412197111<++>Antibiotic->Kindness->Infection
1135413<++>Y<++>0.8892640322015729<++>Differential_(mathematics)->Code->Thoracic_outlet_syndrome
1136769<++>Y<++>0.9991866473974437<++>Lacquer->Brass->Tarnish
118440<++>Y<++>0.7794287084444994<++>Definition->Brooklyn–Manhattan_Transit_Corporation->Medicine
119821<++>Y<++>0.8289089381260273<++>Definition->Curvilinear_coordinates
121171<++>Y<++>0.9236746183603595<++>Definition->Etruscan_civilization
125659<++>Y<++>0.9243819049504125<++>Definition->Preterm_birth
156498<++>Y<++>0.9951922187725896<++>Google_Docs->Autosave
166046<++>Y<++>0.9722765437113997<++>Ethambutol->Therapy->Osteomyelitis
169208<++>Y<++>0.9763904984081142<++>Mississippi->Income_tax
174463<++>Y<++>0.9444240844737418<++>Dog_Day_Afternoon->Dog->Semantics
197312<++>Y<++>0.8524243580136197<++>Group_(mathematics)->Main_Page->Policy
206106<++>Y<++>0.9984726513911077<++>Hotel->St._Louis->Area
227873<++>Y<++>0.9538618238444815<++>Human_body->Redox->Alcohol->Elimination_reaction
246883<++>Y<++>0.7046466212361978<++>Vowel_length->Tick->Survival_skills->Television_presenter
26703<++>Y<++>0.695505587080839<++>United_States_Army->Online_dating_service
273695<++>Y<++>0.7994940831293179<++>Vowel_length->Methadone->Stay_of_execution->System
302846<++>Y<++>0.9999999291388452<++>Caffeine->Twinings->Green_tea
330501<++>Y<++>0.804146658783798<++>Weight->United_States_Postal_Service->Letter_(alphabet)
330975<++>Y<++>0.996579761713109<++>Cost->Installation_(computer_programs)->Wind_turbine
3505<++>Y<++>0.9982497117674316<++>Cardiac_surgery
384356<++>Y<++>0.9944998817120446<++>Uninstaller->Xbox->Windows_10
390360<++>Y<++>0.9763556815701261<++>Ia_(cuneiform)->Suffix->Semantics
405163<++>Y<++>0.9987909780589439<++>Caffeine->Narcotic
42255<++>Y<++>0.8455330333932864<++>Average->Salary->Dental_hygienist->Nebraska
425632<++>Y<++>0.9660983982241111<++>Splitboard->Skiing
426175<++>Y<++>0.9994011991734015<++>Duodenum->Muscle
42752<++>Y<++>0.8009935481520076<++>Average->Salary->Canada->1985
444389<++>Y<++>0.9939103674271949<++>Magnesium->Definition->Chemistry
449367<++>Y<++>0.7916624997735973<++>Semantics->Tattoo->Human_eye
452915<++>Y<++>0.9822391456815329<++>Metabolic_disorder->Medical_sign->Symptom
47210<++>Y<++>0.7671118604971021<++>Weighted_arithmetic_mean->Wedding_dress->Metasomatism->Cost
482726<++>Y<++>0.7674487370523141<++>Projective_variety->Definition
48792<++>Y<++>0.8389210174021245<++>Barclays->Financial_Conduct_Authority->Number
519025<++>Y<++>0.9480360316344636<++>Symptom->Shingles
537060<++>Y<++>0.7097385940332386<++>Village->Frederick_Russell_Burnham
545355<++>Y<++>0.9951076974746371<++>Weather->Novi_Sad
583468<++>Y<++>0.999999934910678<++>Carvedilol
655526<++>Y<++>0.9330128162719924<++>Ezetimibe->Therapy
655914<++>Y<++>0.678786735195569<++>Drive_theory->Poaching
673670<++>Y<++>0.9999833643179875<++>Alpine_transhumance
701453<++>Y<++>0.9643977768213703<++>Statute->Deed
703782<++>Y<++>0.7785827479942069<++>Anterior_cruciate_ligament_injury->Compact_disc
708979<++>Y<++>0.8104485049064436<++>Riding_aids->HIV
730539<++>Y<++>0.8891690146753408<++>Marine_chronometer->Invention
735922<++>Y<++>0.7026236168739335<++>Wool_classing->Petroleum
768208<++>Y<++>0.9013453970344043<++>Pouteria_sapota
779302<++>Y<++>0.9969106553448834<++>Onboarding->Credit_union
794223<++>Y<++>0.9800531144849391<++>Science->Definition->Cytoplasm
```

```
794429<++>Y<++>0.8861366041014064<++>Sculpture->Shape->Space
801118<++>Y<++>1.0<++>Supplemental_Security_Income
804066<++>Y<++>0.9977152604583308<++>Actor->Color
814183<++>Y<++>0.9804059188957711<++>Bit_rate->Standard-definition_television
819983<++>Y<++>0.999999999924519<++>Electric_field
849550<++>Y<++>0.9907674444891965<++>Symptom->Croup
850358<++>Y<++>0.9810309510883796<++>Temperature->Venice->Floruit
914916<++>Y<++>0.7560312455207127<++>Type_species->Epithelium->Bronchiole
91576<++>Y<++>0.8914307302033609<++>Chicken->Food->Wikipedia
945835<++>Y<++>0.7647917173318495<++>Ace_Hardware->Open_set
978031<++>Y<++>0.9999876622967928<++>Berlin_Center,_Ohio
985594<++>Y<++>0.9120158209114781<++>Cambodia
99005<++>Y<++>0.8271551120333056<++>Religious_conversion->Quadraphonic_sound->Metre->Quadraphonic_sound->Inch
999466<++>Y<++>0.9999990098099194<++>Velbert
```

CURRICULUM VITAE

ORCID ID: 0000-0001-9641-9698

Name Surname **: Mohamed Lemine SIDI**

Foreign Language **: Arabic, English, and French.**

Place and Year of Birth **:**

Email **:**

Education and Professional Background:

- 2014-2023, PhD in Computer Engineering, Eskisehir Technical University, Engineering Faculty, Department of Computer Engineering.
- 2009-2011, Master’s degree in Computer Science, Gaston Berger University (Saint-Louis, Senegal), Science Faculty, Department of Computer Science.
- 2005-2009, Bachelor's degree in Computer Science, Gaston Berger University (Saint-Louis, Senegal), Science Faculty, Department of Computer Science.
- 2011-2013, IT Engineer, Mauritanian Audit Court, IT.

Publications and/or Scientific/Artistic Activities:

- Sidi ML, Gunal S. A Purely Entity-Based Semantic Search Approach for Document Retrieval. Applied Sciences. 2023; 13(18):10285. https://doi.org/10.3390/app131810285