UlyssesNERQ: Expanding Queries from Brazilian Portuguese Legislative Documents through Named Entity Recognition

Hidelberg O. Albuquerque[1,2] (✉) **Ellen Souza**[1,3] **Tainan Silva**[1] **Rafael P. Gouveia**[3]
Flavio Junior[1] **Douglas Vitório**[1,2] **Nádia F. F. da Silva**[3,4]
André C.P.L.F. de Carvalho[4] **Adriano L.I. Oliveira**[2]
Francisco Edmundo de Andrade[5]
[1]MiningBR Research Group, Federal Rural University of Pernambuco, Recife, Brazil
[2]Centro de Informática, Federal University of Pernambuco, Recife, Brazil
[3]Institute of Informatics, Federal University of Goiás, Goiás, Brazil
[4]Institute of Mathematics and Computer Sciences, University of São Paulo, São Paulo, Brazil
[5]Brazilian Chamber of Deputies, Brasília, Brazil
{hidelberg.albuquerque,ellen.ramos,flavio.rocha}@ufrpe.br
{tainan206,rafael.p.gouveia2}@gmail.com,
{damsv,alio}@cin.ufpe.br, nadia.felix@ufg.br,
andre@icmc.usp.br, francisco.edmundo@camara.leg.br

Abstract

This study presents UlyssesNERQ, a system designed to improve Information Retrieval for Brazilian Portuguese legislative documents. It uses Named Entity Recognition in Query (NERQ) to expand the queries, seeking to improve an integrated Information Retrieval system. In this sense, a proposal is made to update an existing pipeline, which was evaluated using an experimental approach, with different combinations of text pre-processing techniques, and the use of learning models. Two named entities from a legislative corpus for NER were used. The results show that the combination of RM3 and our method using a BERT model tuned for NER in the legislative domain obtained the best performance, significantly enhancing the accuracy of document retrieval, with an average improvement of around 1.94% (best results) and 8.58% (overall). Additionally, the recall in 20 documents (R@20) has been increased from 0.7356 to 0.7458.

1 Introduction

In Information Retrieval (IR) systems, a query represents a question or a set of keywords entered by the user with the aim of finding specific information. User queries are primarily processed using indexes and ontologies, which rely on exact matches and are not directly visible to users (Azad and Deepak, 2019). In the context of IR, exact match means that terms in the documents and terms in the query must match exactly in order to contribute to a relevance score. One notable method utilizing this approach is Okapi BM25 (Robertson et al., 1994; Robertson and Zaragoza, 2009), which continues to serve as a foundation for many text ranking techniques in both academic research and software industry (Yates et al., 2021). When the terms used by users in their queries do not match the terms used in the search index, it creates a problem known as "term mismatch", which is also referred to as the vocabulary problem (Azad and Deepak, 2019). To tackle this issue, numerous techniques have been proposed, with the majority of them focusing on expanding the initial query by incorporating additional related terms, a process known as Query Expansion (QE) (Azad and Deepak, 2019).

QE is a technique employed to enhance the search results, aiming to make them more precise or comprehensive, and is applied when the initial search results do not meet the user's expectations. This technique seeks to increase the effectiveness of the search by including similar terms in the original query, thereby enabling the retrieval of more relevant documents while reducing the number of irrelevant documents (Zheng et al., 2020; Silva et al., 2021). The initial concept of QE revolves around incorporating user feedback into the retrieval process to enhance the final search results (Rocchio, 1971). Recently, Named Entity Recognition (NER) has been exploit to identify entities in queries and expand them with information from corpora (Lizarralde et al., 2019).

NER is a task in Natural Language Processing (NLP) with applications embracing information extraction, text understanding, and IR. It involves identifying mentions of predefined semantic types or categories within text, such as people, locations,

and organizations (Li et al., 2020). NER is present in multiple areas, such as financial, journalistic, medical/clinical, and in the legal and legislative domains. The number of IR systems that have been using NER has been growing in recent years (Li et al., 2020; Albuquerque et al., 2023a). In this sense, Named Entity Recognition in Query (NERQ) optimizes IR systems by identifying named entities in search strings (Guo et al., 2009). These entities are used to semantically expand the original query, adding or removing candidate entities (Catacora et al., 2022; Khader and Ensan, 2023).

In addition to NER, several classical techniques/resources are cited in literature to expand queries (Azad and Deepak, 2019), e.g, Synonyms detection (Mandal et al., 2019), Relevance Feedback and/or Pseudo-Relevance Feedback (AL-Masri et al., 2016; Vitório et al., 2023), Ontologies (Nevěřilová and Kvaššay, 2018), Thesauruses (Amalia et al., 2021), and Relevance Model 3 (RM3) (Nogueira et al., 2019; Catacora et al., 2022).

In this paper, we introduce *UlyssesNERQ*, an approach that enhances queries by incorporating relevant information for the identified entities. The UlyssesNER-Br (Albuquerque et al., 2022), a corpus of Brazilian legislative documents for NER, was used to identify the entities present in the queries. This research is conducted in the context of the *Ulysses* project, an institutional set of artificial intelligence initiatives with the purpose of increasing transparency, improving the Brazilian Chamber of Deputies' relationship with citizens, and supporting the legislative activity with complex analysis (Almeida, 2021).

This paper is organized as follows: Sec. 2 presents the major related studies. Sec. 3 presents the UlyssesNERQ approach. Sec. 4 details the proposed pipeline and the method used to evaluate the query expansion techniques. Sec. 5 presents and discusses the obtained results. Sec. 6 brings the conclusion and highlights future works.

2 Related Work

Catacora et al. (2022) proposes a legal Information Retrieval system with entity-based query expansion to improve document retrieval in traffic accident litigation. Their system leverages a knowledge base of legal documents and semantic indexes (SAIJ documents and SAIJ thesaurus) to suggest semantically relevant terms related to the user's initial query. Two unsupervised search algorithms, Relevance Model with Entities (RE) and Iterative Relevance Model with Entities (IRE), were implemented for Query Expansion, combined with semantic expansion techniques (MLM and PRMS) and traditional models (TF-IDF and RM3). Results showed that PRMS-based models outperformed MLM-based models, particularly using Mean Average Precision (MAP). Additionally, the automatic expansion model (RE) performed more reliably in interactive entity selection. However, it is important to note that user-selected entities did not consistently lead to better query expansion terms. Overall, this research demonstrates the potential of semantic search systems with QE in the domain.

Silva et al. (2021) proposes a QE technique for Information Retrieval in precision medicine. Their technique uses Multinomial Naïve Bayes (MNB) to extract relevant terms from retrieved documents and combine them with the original query terms to create an expanded query. The method begins with the standard IR process, including text preprocessing and document indexing. Named entities such as disease names and gene variants are then identified and used to construct a "combined query" (CQ). Finally, new terms are extracted using the MNB algorithm and combined with the CQ terms to form the final expanded query. The performance of the QE technique was evaluated using the Clinical Trial corpus, which contains clinical documents, topics, and relevance judgments given by specialists. Results showed a significant improvement in document retrieval performance with QE, with MAP increasing by approximately 30%, which suggests that MNB-based query expansion can significantly enhance the precision of document retrieval.

Kandasamy and Cherukuri (2020) created a method for Named Entity Disambiguation to improve QE in question-answering (QA) systems for general domain. The study uses an adapted Lesk similarity measure (commonly used for word sense disambiguation) to identify and expand the most relevant named entities in the query. The proposed method was evaluated using two versions of a dataset of questions (TREC QA dataset), incorporating Wikipedia articles and disambiguation pages, and comparing it to the state-of-the-art. The results showed an enhancement in the accuracy of QA systems by expanding queries with relevant entities, reporting higher precision and recall averages compared to the state-of-the-art, smoothly

Study	Domain	Data source	Algorithms/Models	Techniques	Metrics
Catacora et al. (2022)	Traffic litigation	SAIJ documents and thesaurus	RE, IRE	NER, MLM, PRMS, TF-IDF, RM3	MAP
Silva et al. (2021)	Precision medicine	Clinical Trial corpus	MNB	NER, Text preprocessing, Document indexing, Combined query (CQ)	MAP
Kandasamy and Cherukuri (2020)	General	Wikipedia, TREC QA dataset	Adapted Lesk similarity measure	NER, Selecting keyword meanings, Subject area identification	Precision, Recall, and F1-score
Sarwat et al. (2019)	General	TREC QA dataset	Topical and functional similarities	NER, SQE, and TQE	Precision, Recall[1], and MAP
Tang et al. (2015)	General	Chinese Wikipedia and a Knowledge base	Re-ranking, Word Embedding Similarity and Entity Frequency	NER, use of Search Engine (Baidu), Filter Rules, Synonyms	Precision, Recall, and F1-score
Our proposal	Legislative	UlyssesNER-BR corpus	CRF, BERT, and Bertikal	NER, Text preprocessing, RM3, and Synonym	Recall at 20 (R@20)

Table 1: Comparison of related works.[1]In terms at 10 and 20 documents retrieval.

overcoming in disambiguating organization and miscellaneous-type entity mentions. Overall, the method holds the potential to enhance QA systems performance.

Sarwar et al. (2019) proposed a retrieval approach using a single training sentence to extract more data for information extraction tasks. It aims to retrieve sentences with relevant and novel entities of the same type. Topical and functional similarities are used to rank candidate sentences, captured through sentence embedding (SQE and TQE), while QE broadens the training sentence representation. Functional similarity is achieved by examining NER tags in candidate sentences. Evaluation on a dataset of 120 list questions from TREC List QA datasets shows that the proposed approach outperforms the baseline BM25 ranking algorithm, with significant improvements in precision, recall, and MAP. The approach also retrieves a high percentage of target entities in top-ranked sentences. It holds promise in addressing data sparsity in information extraction tasks.

Tang et al. (2015) presented a system for NER and linking in search queries, addressing challenges like short context, nonstandard text, and diverse entity representations. The system employs a rule-based approach for NER, generating candidate entities using a search engine and Wikipedia, and applies a re-ranking method to score and obtain linking results. The techniques utilized include rule-based entity recognition, a search engine and Wikipedia-based candidate entity generation, and a hybrid re-ranking method using textual and semantic matching, word embedding similarity, and entity frequency. The results pointed out an average F1 score of $\sim$0.61. It outperforms the third-ranked system in terms of link-recall and link-F1 but falls behind the top-ranked system in terms of link-precision and average F1.

This study enhances Information Retrieval in the Brazilian legislative domain by applying advanced NLP techniques such as NER, Synonym Detection, and RM3, improving the search and retrieval of relevant documents. The effectiveness of these models is validated using recall for the top 20 documents, aiming to boost precision and recall for legal professionals, scholars, and the public seeking legislative information. The Table 1 shows distinctions between the selected studies and our proposal. It is important to note that studies cannot be directly equated, as they belong to different domains and involve variances in models, techniques, and metrics.

3 UlyssesNERQ

UlyssesNERQ is a NER system that employs the NERQ technique for query expansion and operates within the context of the Brazilian legislative domain. In this context, the queries submitted by parliamentarians to the Brazilian Legislative Consultation Department (*Conle*)[1] aim to retrieve bills and other legislative consultations. This method works together with another internal IR system used by Conle (Souza et al., 2021b) (Figure 1(A)), acting on the extension of the pre-processed consultations, performing NER tasks to expand the original query (Figure 1(B)).

UlyssesNERQ is able to identify 18 types of named entities found within the UlyssesNER-Br Corpus (Albuquerque et al., 2022), a corpus of Brazilian legislative documents for NER (Table 2). This includes entities related to bills (*FUNDprojetolei*) and legislative consultations (*FUNDsolicitacaotrabalho*), among others. The choice to use only two types of entities was guided by internal prerogatives of the legislative project, highlighting the critical importance of these entities in representing a wide spectrum of legislative activities used. Other factors that influenced this decision in-

[1]https://www2.camara.leg.br/a-camara/estruturaadm/consultoria-geral/consultoria-legislativa

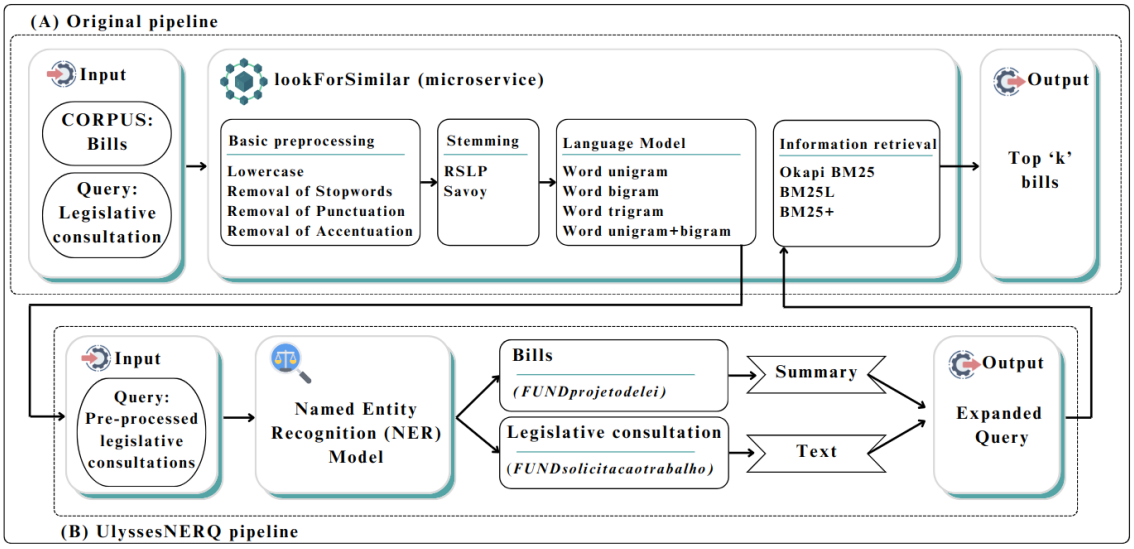


Figure 1: (A) Brazilian Chamber of Deputies' Information Retrieval pipeline (Souza et al., 2021b). (B) UlyssesNERQ pipeline.

Category	Type	Description	Example
DATA (Date)	—	Date	01 de janeiro de 2020
EVENTO (Event)	—	Event	Eleições de 2018
FUNDAMENTO (Law foundation)	FUNDlei	Legal norm	Lei no 8.666, de 21 de junho de 1993
	FUNDapelido	Legal norm nickname	Estatuto da Pessoa com Deficiência
	FUNDprojetodelei	Bill	PEC 187/2016
	FUNDsolicitacaotrabalho	Legislative consultation	Solicitação de Trabalho n^o 3543/2019
LOCAL (Location)	LOCALconcreto	Concrete place	Niterói-RJ
	LOCALvirtual	Virtual place	Jornal de Notícias
ORGANIZAÇÃO (Organization)	ORGpartido	Political party	PSB
	ORGgovernamental	Governamental organization	Câmara do Deputados
	ORGnãogovernamental	Non-governmental organization	Conselho Reg. de Medicina (CRM)
PESSOA (Person)	PESSOAindividual	Individual	Jorge Sampaio
	PESSOAgrupoind	Group of individuals	Família Setúbal
	PESSOAcargo	Occupation	Deputado
	PESSOAgrupocargo	Group of occupations	Parlamentares
PRODUTO DE LEI (Law product)	PRODUTOsistema	System product	Sistema Único de Saúde (SUS)
	PRODUTOprograma	Program product	Programa Minha Casa, Minha Vida
	PRODUTOoutros	Others products	Fundo partidário

Table 2: UlyssesNER-Br corpus: categories and types (Albuquerque et al., 2022).

Originated bill	Legislative consultation (user's query)	Entities	Expanded query
(A) PL XXXX/2019	*Criação de PL, com base nos dois esboços encaminhados anexo* (Make of bill based on the two sketches sent in the attachment)	0	*Criação de PL, com base nos dois esboços encaminhados anexo*
(B) PL XXXX/2019	*Complementar parecer em função da apensação do PL XXXX/19 ao mesmo* (Complementary opinion according to the PL XXXX/19)	1	*Complementar parecer em função da apensação do PL XXXX/19 ao mesmo* ***Altera o art. X da Lei n. XX/XXXX, para modificar a sua cláusula de vigência*** (Amends art. Xrd of Law no. XX/XXXX, to modify its validity clause)
(C) PL XXXX/2019	*Parlamentar solicita aprovação* (Parliamentarian requests approval)	0	*Parlamentar solicita aprovação* ***relatoria aprovação solicita pela parlamentar emenda aval 2022 origin*** (reporting approval requested for the parliamentary amendment approval 2022 origin)
(D) PL XXXX/2019	*Projeto para restabelecer na CLT a proibição de terceirização para atividade fim* (Project to prohibit the outsourcing of core activity in the CLT)	0	*Projeto para restabelecer na CLT a proibição de terceirização para atividade fim* ***projecto ideia atividade programa proibições ocupação ato interdição proibição*** (project idea activity program prohibitions occupation act interdiction prohibition)

Table 3: Anonymized samples of legislative consultations (Souza et al. (2021b), adapted). Query expansions in bold.

cluded considerations such as model training time and computational resources employed, limitations that will be subject to analysis later.

The process begins with the insertion of a search string with a legislative consultation that goes through several text pre-processing steps. After that, the NER model is applied to identify the entities present in the request, searching the database for previous bills or legislative consultations. If a document is found, the query is enhanced by adding text from the summary section (for bills) or the content of the legislative consultation itself. This procedure ends with an expanded query enriched with a broader set of terms, thus improving the effectiveness of the IR system. The new query is then used in the original system structure. If no

previous bill or request is found in the database, the original query is used (see Table 3, (A) and (B)).

To choose the NER model used, three state-of-the-art models were assessed, one of them based on the CRF model (Albuquerque et al., 2022), other based on the BERT model (Albuquerque et al., 2023b), and other called *BERTikal* (Polo et al., 2021), also based on BERT. When necessary, the model chosen was pre-trained to the legislative domain using the corpus selected. Furthermore, combinations of the chosen model with other NLP techniques were evaluated.

4 Method

4.1 Brazilian Chamber of Deputies' pipeline

As mentioned, UlyssesNERQ updates the IR pipeline proposed by Souza et al. (2021b). In their paper, the authors evaluated three BM25 algorithms and 21 combinations of pre-processing techniques to decide which configuration was the most suitable for the retrieval of legislative documents in the Conle's scenario. Figure 1(A) presents the pipeline evaluated by them.

The best configuration used the BM25L algorithm together with texts pre-processed with lowercase, the removal of punctuation, accentuation, and stopwords, a combination of the unigram and bigram language models, and the stemmer Savoy (Savoy, 2006). The Stemming algorithm choice was also confirmed by a later work (Souza et al., 2021a), in which the impact of Stemming in this scenario was evaluated, concluding that Savoy was the best to be used with BM25L. Thus, this is the IR model currently used by Conle.

4.2 Query Expansion techniques

In addition to UlyssesNERQ, were evaluated other techniques for Query Expansion: using Relevance Model 3 (RM3) and using Synonyms detection.

4.2.1 Relevance Model 3

RM3 uses the set of retrieved documents for the initial query to create a relevance model. The query is, then, expanded using relevant terms from the retrieved documents. This technique aims to improve the performance of the retrieval process, especially when the initial query is vague (see Table 3 (C)). It is usually applied in search engines and IR systems to improve the relevance level of the retrieved documents (Nogueira et al., 2019).

4.2.2 Query Expansion with Synonyms

The use of synonyms for QE involves expanding the initial query through the addition of synonyms and similar phrases, i.e., terms with similar meanings to those used in the original query. This technique extends the spectrum of related terms, improving the chance to retrieve pertinent information (see Table 3 (D)). It is also more useful when the initial query is vague or in cases in which different words can describe the same concept (Mandal et al., 2019).

4.3 Corpora

Two corpora were used, containing bills and legislative consultations. These corpora are part of a larger NER dataset comprised of legislative documents from the Brazilian Chamber of Deputies (Albuquerque et al., 2022). This larger dataset contains 11 types of entities based on HAREM (Santos and Cardoso, 2006) grouped into 7 categories, and 7 legal entities grouped into 2 categories. Only the Bill corpus is publicly available[2] , while the Legislative consultations corpus contains confidential information and cannot be shared. As mentioned, all the named entities are shown in Table 2.

The Bill corpus used in this study comprises 57,109 publicly available legislative proposals, encompassing the three most common types: Law Project (*Projeto de Lei - PL*), Complementary Law Project (*Projeto de Lei Complementar - PLC*), and Constitutional Amendment Proposal (*Proposta de Emenda à Constituição - PEC*). This dataset represents an updated version of the corpus used in Souza et al. (2021b), which only included bills up to the year 2020.

The Legislative Consultations corpus employed in this study corresponds to the same dataset used by Souza et al. (2021b). This corpus was curated by the Conle department of the Chamber of Deputies and leverages the IR model to retrieve bills and other legislative consultations based on requests from parliamentarians. Following the retrieval, parliamentarians employ this information to formulate new bills for consideration in the House. The dataset used in this research comprises 295 anonymized requests, along with the respective names of the formulated bills, as depicted in Table 3. Notably, any data which may allow identifying the parliamentarian who made the request to

[2]https://github.com/Convenio-Camara-dos-Deputados/ulyssesner-br-propor

Config.[1]	BM25L[2]	UlyssesNERQ CRF[3]	UlyssesNERQ BERTikal[4]	UlyssesNERQ BERT[5]	RM3[6]	Synonym[7]
basic preprocessing						
1	0.6576	**0.6780**	0.6746	**0.6780**	0.6610	0.6475
2	0.6847	**0.7085**	**0.7085**	**0.7085**	0.6712	0.6780
3	0.7186	**0.7288**	0.7254	**0.7288**	0.6949	0.6780
4	0.7254	0.7356	0.7356	**0.7390**	0.7051	0.6780
5	0.7153	**0.7254**	0.7220	**0.7254**	0.7085	0.6814
stemming						
6	0.6678	0.6881	0.6847	0.6881	**0.6949**	0.6542
7	0.6508	0.6712	0.6678	0.6712	**0.6814**	0.6441
8-4	0.7288	0.7322	0.7288	**0.7356**	0.7186	0.6881
9-4	0.7220	0.7254	0.7220	**0.7288**	0.7085	0.6814
8	0.7220	0.7288	0.7254	**0.7322**	0.7220	0.6881
9	0.7220	0.7288	0.7254	**0.7322**	0.7085	0.6881
word n-gram						
10	0.5966	0.6068	0.6746	**0.6780**	0.5864	0.6068
11	0.4780	**0.5119**	0.5085	0.5085	0.4881	**0.5119**
12	0.6610	**0.6746**	**0.6746**	**0.6746**	**0.6746**	0.6678
word n-gram + basic preprocessing						
13-4	0.6136	0.6271	0.6237	0.6271	0.6203	**0.6305**
14-4	0.5119	**0.5322**	0.5288	**0.5322**	0.5153	**0.5322**
15-4	0.6949	0.5322	0.5288	0.5322	0.5153	**0.7017**
13	0.6000	**0.6169**	**0.6169**	**0.6169**	0.5932	**0.6169**
14	0.4712	**0.4949**	0.4915	0.4915	0.4712	**0.4949**
15	0.6983	0.7051	0.7051	0.7051	0.7051	**0.7119**
word n-gram + basic preprocessing + RSLP						
16-4	0.6441	**0.6542**	0.6508	**0.6542**	0.6475	**0.6542**
17-4	0.5356	**0.5593**	0.5559	**0.5593**	0.5458	**0.5593**
18-4	0.7186	**0.7288**	0.7220	0.7254	0.7186	**0.7288**
16	0.6373	0.6441	0.6407	**0.6475**	0.6373	0.6441
17	0.4847	**0.5051**	0.5017	**0.5051**	0.4949	**0.5051**
18	0.7356	0.7424	0.7424	**0.7458**	0.7356	0.7356
word n-gram + basic preprocessing + Savoy						
19-4	0.6407	**0.6542**	0.6508	**0.6542**	0.6441	**0.6542**
20-4	0.5254	**0.5492**	0.5458	**0.5492**	0.5288	**0.5492**
21-4	0.7085	**0.7186**	0.7153	**0.7186**	0.6983	0.7085
19	0.6305	0.6373	0.6339	0.6407	**0.6441**	0.6407
20	0.4814	**0.5017**	0.4983	**0.5017**	0.4847	**0.5017**
21	0.7153	0.7186	0.7186	**0.7254**	0.7186	**0.7254**

Table 4: Analysis of query expansion techniques individually with recall for 20 documents. [1]Configuration of technique combination. Implementations by: [2](Souza et al., 2021b), [3](Albuquerque et al., 2022), [4](Polo et al., 2021), [5](Albuquerque et al., 2023b), [6](Nogueira et al., 2019), [7](Azad and Deepak, 2019).

Conle has been omitted for privacy reasons.

4.4 Experimental Configuration

The QE models were assessed in the same 21 configurations from Souza et al. (2021b), built combining the pre-processing techniques shown in Figure 1(A). However, in our experiments, was observed that the configuration 4 (lowercase + punctuation and accentuation removal) outperformed the configuration 5 (lowercase + punctuation, accentuation, and stopword removal). For this reason, we included 11 more experiments, combining other techniques with configuration 4. And, as it obtained the best results, the BM25L was chosen as the IR algorithm, thus we did not perform experiments with the BM25 Okapi and Plus variants. As baseline, we consider the system without QE and using BM25L.

As each query have only one relevant document (Table 3), the results were evaluated in terms of recall at 20 (*R@20*), which corresponds to the fraction of relevant documents that were retrieved among the top 20 retrieved documents. The decision to use the R@20 metric was based on emphasizing comprehensive coverage of relevant documents, prioritizing the identification of top items of interest. Besides, it aims to underscore the importance of finding highly relevant documents in the project's context.

5 Results and Discussion

5.1 Individual Results

Comparing the results individually (Table 4), it can see that the model using UlyssesNERQ with BERT obtained the best individual result with configura-

Config.[1]	UlyssesNERQ BERT	UlyssesNERQ BERT + RM3	UlyssesNERQ BERT + Synonyms	RM3 + UlyssesNERQ BERT	RM3 + UlyssesNERQ BERT + Synonyms
basic preprocessing					
1	0.6780	0.6814	0.6475	**0.6881**	0.6508
2	**0.7085**	0.6915	0.6814	0.6983	0.6678
3	**0.7288**	0.7051	0.6746	0.7119	0.6746
4	**0.7390**	0.7119	0.6780	0.7186	0.6746
5	**0.7254**	0.7153	0.6780	0.7220	0.6712
stemming					
6	0.6881	0.7119	0.6542	**0.7186**	0.6881
7	0.6712	0.7017	0.6475	**0.7085**	0.6576
8-4	**0.7356**	0.7254	0.7085	0.7322	0.7186
9-4	**0.7288**	0.7153	0.6949	0.7220	0.7017
8	0.7322	0.7288	0.7119	**0.7356**	0.7153
9	**0.7322**	0.7153	0.7017	0.7220	0.7017
word n-gram					
10	**0.6780**	0.6068	0.6034	0.6102	0.6102
11	0.5085	0.5085	0.5085	**0.5119**	**0.5119**
12	0.6746	0.6780	0.6678	**0.6814**	0.6678
word n-gram + basic preprocessing					
13-4	0.6271	0.6339	0.6271	0.6373	**0.6407**
14-4	0.5322	0.5356	0.5322	**0.5390**	0.5356
15-4	0.5322	0.5356	**0.6983**	0.5390	0.5356
13	**0.6169**	0.6068	0.6102	0.6102	0.6068
14	0.4915	0.4915	0.4915	**0.4949**	**0.4949**
15	0.7051	**0.7153**	0.7119	**0.7153**	0,7051
word n-gram + basic preprocessing + RSLP					
16-4	0.6542	0.6542	0.6542	**0.6610**	**0.6610**
17-4	0.5593	0.5695	0.5593	**0.5729**	0.5695
18-4	0.7254	0.7220	0.7254	0.7288	**0.7322**
16	0.6475	0.6475	0.6441	**0.6508**	0.6441
17	0.5051	0.5153	0.5051	**0.5186**	0.5153
18	**0.7458**	0.7424	**0.7458**	**0.7458**	**0.7458**
word n-gram + basic preprocessing + Savoy					
19-4	0.6542	0.6542	0.6508	**0.6610**	0.6576
20-4	0.5492	0.5525	0.5492	**0.5559**	0.5525
21-4	**0.7186**	0.7017	0.7051	0.7085	**0.7186**
19	0.6407	0.6508	0.6373	**0.6542**	0.6441
20	0.5017	0.5051	0.5017	**0.5085**	0.5051
21	0.7254	0.7288	**0.7390**	0.7288	0.7288

Table 5: Analysis of the combination of query expansion techniques, with recall for 20 documents. [1]Configuration of technique combination.

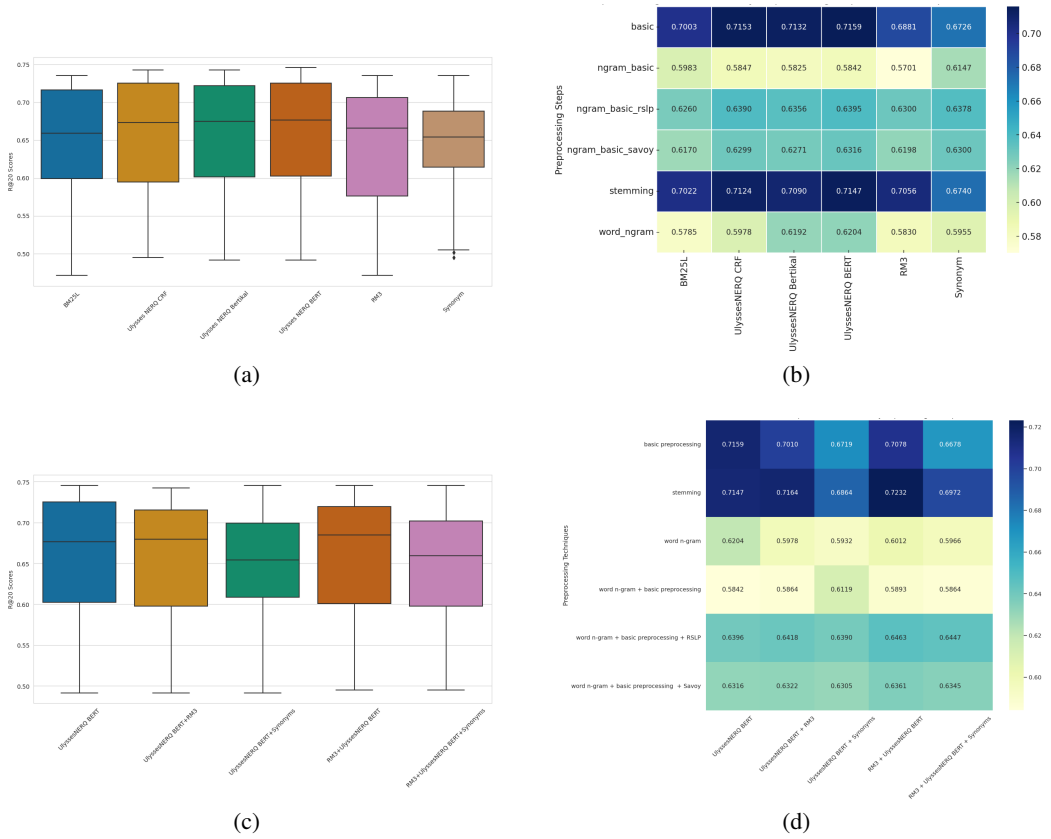
tion of techniques combination number 18 (Config. n.18). However, seeking a statistical basis to justify this choice, the calculation of the mean and standard deviation was used for each technique, followed by the application of statistical tests, as will be presented below. These results are shown in Table 6(A).

The Shapiro-Wilk test (Shapiro and Wilk, 1965) was initially applied to assess the normality of the distributions of results generated by each technique, which is crucial as many statistical techniques assume normality in the data. If the distributions do not follow a normal distribution, as indicated by the Shapiro-Wilk test results (p-value < 0.05), it necessitates the use of non-parametric statistical approaches. Since the normality assumption was not met, the Kruskal-Wallis non-parametric test (Dodge, 2008) was employed as a robust alternative to compare medians between techniques and identify statistically significant differences in results between multiple groups. The test revealed no statistically significant differences (p-value > 0.05), suggesting similar performances among the query expansion techniques. Additionally, confidence intervals for the models were calculated, showing overlapping 95% confidence intervals (0.6094 to 0.6822). Thus, these analyses did not yield a statistically supported conclusion about the superior technique.

Evaluating the best overall performance, was examined the average recall scores during preprocessing and identified the top-performing approach. Figure 2(A) and (B) illustrates that UlyssesNERQ BERT consistently achieves high recall scores across various pre-processing configurations, aligning with our earlier findings. Fur-

Tecnhique	Mean ± Standart deviation	Shapiro-Wilk (p-value)	Kruskal-Wallis (p-value)	Confidence intervals
(A) Individual results				
BM25L	0.6406±0.0864	∼0.00065	∼0.729	0.6094 to 0.6718
UlyssesNERQ CRF	0.6489±0.0831	∼0.00095		0.6189 to 0.6789
UlyssesNERQ BERTikal	0.6484±0.0832	∼0.00054		0.6184 to 0.6784
UlyssesNERQ BERT	**0.6519±0.0840**	∼0.00064		0.6216 to 0.6822
RM3	0.6357±0.0851	∼0.00070		0.6050 to 0.6664
Synonyms	0.6403±0.0712	∼0.00268		0.6146 to 0.6660
(B) Combined models results				
Ulysses NERQ BERT	0.6519±0.0840	∼0.0006	∼0.7155	0.6216 to 0.6822
Ulysses NERQ BERT+RM3	0.6487±0.0801	∼0.0007		0.6198 to 0.6776
Ulysses NERQ BERT+Synonyms	0.6421±0.0741	∼0.0071		0.6154 to 0.6688
RM3+Ulysses NERQ BERT	**0.6535±0.0810**	∼0.0006		0.6243 to 0.6827
RM3+Ulysses NERQ BERT+Synonyms	0.6408±0.0748	∼0.0081		0.6138 to 0.6678

Table 6: Statistical tests to (A) individual and (B) combination results.



(a) (b) (c) (d)

Figure 2: Average Recall by results: (a)-(b) individual, and (c)-(d) combined techniques.

thermore, the results reveal a higher level of consistency and superior performance in various phases (outperforming in 23/32 configurations) while maintaining an acceptable standard deviation level. This confirms the individual model's effectiveness, and the Config. n.18 as our choice.

From a total of 295 queries, the UlyssesNERQ BERT model expanded 33 of them, which contained at least one of the two used entities. The *FUNDprojetodelei* entity appeared in 41 queries, however, it was incorrectly identified in eight queries. Also, two *FUNDprojetodelei* entities were not found in the corpus. The *FUNDsolicitacaotrabalho* entity appeared in 17 queries, that were not expanded as this corpus has confidential information and it was not available in our experiments.

5.2 Combined Models Results

Due to UlyssesNERQ BERT's best performance, it was combined with RM3 and Synonym techniques, previously tested, to assess their combined efficacy.

From Table 5, it is possible to observe that there

was a tie among the four best individual results, all with configuration no. 18. A statistical analysis of the results was conducted once more to validate which combination had the best result, based on average and standard deviation, demonstrating two best results (UlyssesNERQ BERT and RM3+UlyssesNERQ BERT), with a slight improvement in the latter. The Shapiro-Wilk test results once more confirmed that the data does not conform to a normal distribution. In turn, the Kruskal-Wallis test also indicated that there is no statistically significant difference. Confidence intervals analyzed (Table 6(B)) showed overlapping results at 95% confidence (0.6138 to 0.6827), indicating again that the differences between the groups may be narrow. Minor variations in the intervals from combining models did not constitute significant differences, suggesting no configuration consistently outperformed the others.

Evaluating the best overall performance, we used the same previous method. Figure 2(C) and (D) showed that the combination of RM3 and UlyssesNERQ BERT consistently achieves high recall scores across various pre-processing configurations. In addition, it achieved the highest average recall and an acceptable standard deviation, and increased consistency and superior performance (outperforming in 18/32 configurations), suggesting this model is the most effective.

From a total of 295 queries, the combination RM3 + UlyssesNERQ BERT expanded all of them using the RM3 technique, while at least 30 were extended using NER. The *FUNDprojetodelei* entity appeared in 39 queries, however, it was incorrectly identified in nine. Again, two *FUNDprojetodelei* entities were not found in the corpus, and the *FUNDsolicitacaotrabalho* entity appeared in 17 queries, which were not expanded, as previously explained.

6 Conclusion

This paper describes the UlyssesNERQ system, an update to the IR pipeline employed by the Brazilian Chamber of Deputies for the retrieval of legislative documents. The system implements the Named Entity Recognition in Query (NERQ) approach to detect entities within search queries, subsequently enriching these queries with data from the UlyssesNER-Br corpus, a collection of Brazilian legislative NER documents. The method uses two specific corpora from the Chamber of Deputies, employing two entities, "FUNDprojetodelei" for bills and "FUNDsolicitacaotrabalho" for legislative consultations, to enhance user queries with relevant details.

To validate the pipeline, a variety of configurations were extensively assessed, along with specialized pre-processing techniques as the most effective for IR, considering the BM25L results as a baseline. The Shapiro-Wilk and Kruskal-Wallis statistical tests were performed, which demonstrated that there were no statistically significant differences in choosing the best models (Table 6 and Figure 2). The results were then analyzed considering the best overall average performance, to confirm the initial results. Tables 4 and 5 showed the results, demonstrating that Query Expansion, integrating the RM3 and UlyssesNERQ BERT model, consistently enhances the retrieval performance of BM25L, with an average improvement of about 1.94% (best results) and 8.58% (overall). Moreover, the metric for recall at 20 documents (R@20) was increased from 0.7356 to 0.7458.

Several limitations of our approach have to be listed: the use of only two entities from the legislative corpus, a lack of comparable Portuguese-language studies on legislative QE for benchmarking, and an unexplored area concerning the impact of new generative language models, and query expansion on end-user experience. These limitations point out our future research directions. We plan to enhance the UlyssesNERQ system by incorporating these limitations, mainly using a wider range of entities and employing additional techniques such as a legislative thesaurus to refine query precision and retrieval outcomes. Further, we will extend our experiments to include the latest BERT and Large Language Models.

Acknowledgements

This research is carried out in the context of the Ulysses Project, of the Brazilian Chamber of Deputies. Ellen Souza and Nadia Félix are supported by FAPESP , agreement between USP and the Brazilian Chamber of Deputies. André C.P.L.F. de Carvalho and Adriano L.I. Oliveira are supported by CNPq. To the Brazilian Chamber of Deputies, to the Institute of Artificial Intelligence (IAIA) and to research funding agencies, to which we express our gratitude for supporting the research.

References

Hidelberg O. Albuquerque, Rosimeire Costa, Gabriel Silvestre, Ellen Souza, Nádia F. F. da Silva, Douglas Vitório, Gyovana Moriyama, Lucas Martins, Luiza Soezima, Augusto Nunes, Felipe Siqueira, João P. Tarrega, Joao V. Beinotti, Marcio Dias, Matheus Silva, Miguel Gardini, Vinicius Silva, André C. P. L. F. de Carvalho, and Adriano L. I. Oliveira. 2022. UlyssesNER-Br: A corpus of brazilian legislative documents for named entity recognition. In *Computational Processing of the Portuguese Language*, pages 3–14, Cham. Springer International Publishing.

Hidelberg O Albuquerque, Ellen Souza, Carlos Gomes, Matheus Henrique de C. Pinto, Ricardo P.S. Filho, Rosimeire Costa, Vinícius Teixeira de M. Lopes, Nádia F.F. da Silva, André C.P.L.F. de Carvalho, and Adriano L.I. Oliveira. 2023a. Named entity recognition: a survey for the portuguese language. *Procesamiento del Lenguaje Natural*, 70:171–185.

Hidelberg O. Albuquerque, Ellen Souza, Adriano L. I. Oliveira, David Macêdo, Cleber Zanchettin, Douglas Vitório, Nádia F. F. da Silva, and André C. P. L. F. de Carvalho. 2023b. On the assessment of deep learning models for named entity recognition of brazilian legal documents. In *Progress in Artificial Intelligence*, pages 93–104, Cham. Springer Nature Switzerland.

Mohannad ALMasri, Catherine Berrut, and Jean-Pierre Chevallet. 2016. A comparison of deep learning based query expansion with pseudo-relevance feedback and mutual information. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*, pages 709–715. Springer.

P. G. R. Almeida. 2021. Uma jornada para um Parlamento inteligente: Câmara dos Deputados do Brasil. *Red Información*, 24.

Ivanda Zevi Amalia, Akbar Noto Ponco Bimantoro, Agus Zainal Arifin, Maryamah Faisol, Rarasmaya Indraswari, and Riska Wakhidatus Sholikah. 2021. Indonesian-translated hadith content weighting in pseudo-relevance feedback query expansion. *Jurnal Ilmiah Kursor*, 11(1).

Hiteshwar Kumar Azad and Akshay Deepak. 2019. Query expansion techniques for information retrieval: A survey. *Information Processing Management*, 56(5):1698–1735.

Joel Arnaldo Gimenez Catacora, Ana Casali, and Claudia Deco. 2022. Legal information retrieval system with entity-based query expansion: Case study in traffic accident litigation. *Journal of Computer Science and Technology*, 22(2):e12–e12.

Yadolah Dodge. 2008. *The concise encyclopedia of statistics*. Springer Science & Business Media.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 267–274, New York, NY, USA. Association for Computing Machinery.

Saravanakumar Kandasamy and Aswani Kumar Cherukuri. 2020. Query expansion using named entity disambiguation for a question-answering system. *Concurrency and Computation: Practice and Experience*, 32(4):e5119.

Ayesha Khader and Faezeh Ensan. 2023. Learning to rank query expansion terms for covid-19 scholarly search. *Journal of Biomedical Informatics*, 142:104386.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Ignacio Lizarralde, Cristian Mateos, Juan Manuel Rodriguez, and Alejandro Zunino. 2019. Exploiting named entity recognition for improving syntactic-based web service discovery. *Journal of Information Science*, 45(3):398–415.

Aritra Mandal, Ishita K Khan, and Prathyusha Senthil Kumar. 2019. Query rewriting using automatic synonym extraction for e-commerce search. In *eCOM@ SIGIR*.

Zuzana Nevěřilová and Matej Kvaššay. 2018. Understanding search queries in natural language. In *RASLAN*, pages 85–93. Tribun EU.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

Felipe Maia Polo, Gabriel Caiaffa Floriano Mendonça, Kauê Capellato J. Parreira, Lucka Gianvechio, Peterson Cordeiro, Jonathan Batista Ferreira, Leticia Maria Paz de Lima, Antônio Carlos do Amaral Maia, and Renato Vicente. 2021. Legalnlp – natural language processing methods for the brazilian legal language.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. In *Text Retrieval Conference*.

J. J. Rocchio. 1971. Relevance feedback in information retrieval. In G. Salton, editor, *The Smart retrieval system - experiments in automatic document processing*, pages 313–323. Englewood Cliffs, NJ: Prentice-Hall.

Diana Santos and Nuno Cardoso. 2006. A golden resource for named entity recognition in portuguese. In *Computational Processing of the Portuguese Language*, pages 69–79, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sheikh Muhammad Sarwar, John Foley, Liu Yang, and James Allan. 2019. Sentence retrieval for entity list extraction with a seed, context, and topic. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, page 209–212.

Jacques Savoy. 2006. Light stemming approaches for the french, portuguese, german and hungarian languages. In *Proceedings of the 2006 ACM Symposium on Applied Computing*, SAC '06, page 1031–1035. Association for Computing Machinery.

S. S. Shapiro and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611.

Sergio Silva, Adrián Seara Vieira, Pedro Celard, Eva Lorenzo Iglesias, and Lourdes Borrajo. 2021. A query expansion method using multinomial naive bayes. *Applied Sciences*, 11(21):10284.

Ellen Souza, Gyovana Moriyama, Douglas Vitório, André Carlos Ponce de Leon Ferreira de Carvalho, Nádia Félix, Hidelberg Albuquerque, and Adriano L. I. Oliveira. 2021a. Assessing the impact of stemming algorithms applied to brazilian legislative documents retrieval. In *Proceedings of the 13th Brazilian Symposium in Information and Human Language Technology*, pages 227–236. SBC.

Ellen Souza, Douglas Vitório, Gyovana Moriyama, Luiz Santos, Lucas Martins, Mariana Souza, Márcio Fonseca, Nádia Félix, André Carlos Ponce de Leon Ferreira de Carvalho, Hidelberg O. Albuquerque, and Adriano L. I. Oliveira. 2021b. An information retrieval pipeline for legislative documents from the brazilian chamber of deputies. In *Legal Knowledge and Information Systems*, pages 119–126. IOS Press.

Gongbo Tang, Yuting Guo, Dong Yu, and Endong Xun. 2015. A hybrid re-ranking method for entity recognition and linking in search queries. In *Natural Language Processing and Chinese Computing*, pages 598–605, Cham. Springer International Publishing.

Douglas Vitório, Ellen Souza, Lucas Martins, Nádia FF da Silva, Adriano LI Oliveira, Francisco Edmundo de Andrade, et al. 2023. Building a relevance feedback corpus for legal information retrieval in the real-case scenario of the brazilian chamber of deputies.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: BERT and beyond. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 1–4, Online. Association for Computational Linguistics.

Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, and Andrew Yates. 2020. BERT-QE: Contextualized Query Expansion for Document Re-ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4718–4728, Online. Association for Computational Linguistics.