

Geodata source retrieval in PDOK by multilingual/semantic query expansion

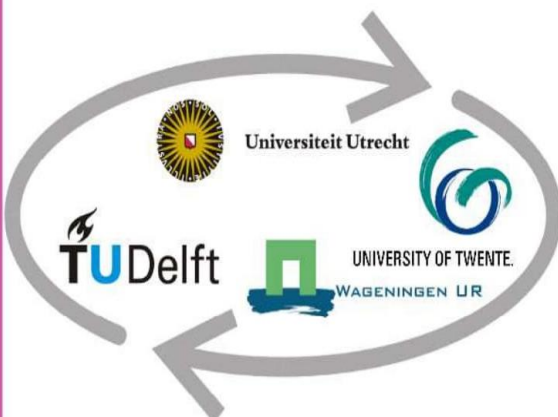
Master Thesis

Maryam Sajjadian

b.sajjadianjaghargh@students.uu.nl

Supervisor: Dr. Simon Scheider

Responsible Professor: Prof. dr. Stan Geertman



Contents

1. Introduction.....	1
1.1. Background.....	1
1.2. Objects and research questions	2
1.3. Scope of research and limitations.....	3
2. literature review	4
2.1. Geo-analytic question corpora and analyzing geographic questions	4
2.2. Query expansion	4
2.2.1. Query expansion using WordNet.....	5
2.2.1.1. Similarity methods.....	5
2.2.1.2. Relatedness between concepts	6
2.2.2. Cross-lingual information retrieval.....	6
2.3. Evaluation methods	7
3. Methodology and implementation	8
3.1. An overview of the methodology	8
3.2. Dataset and infrastructures.....	9
3.3. The first phase: Keyword gathering	10
3.4. The second phase: Answer dataset.....	11
3.5. The third phase: Gold standard	11
3.6. The fourth phase: Defined local scenarios.....	12
3.6.1. Baseline scenario.....	12
3.6.2. Multilingual scenario.....	12
3.6.3. Multilingual WordNet scenario	13
3.6.3.1. Computing synsets.....	13
3.6.3.2. Calculating similarity in WordNet	15
3.6.3.3. Calculating semantic overlay	17
3.7. Scenarios algorithms.....	18
3.8. The fifth phase: IR evaluation metrics	20
3.9. The sixth phase: Online tools scenarios.....	20
4. Results and discussion	23
4.1. Results of the local scenarios	23
4.1.1. Results of the local baseline scenario	23
4.1.1.1. Discussion on the results of the local baseline scenario	24
4.1.2. Results of the local multilingual scenario	24
4.1.2.1. Discussion on results of the local multilingual scenario	25
4.1.3. Results of the local WordNet scenario.....	26
4.1.3.1. Discussion on results of the local WordNet scenario	27
4.2. Final results of the local scenarios.....	28
4.2.1. Discussion on final results of the local scenarios.....	28
4.3. Results of the online tools scenarios.....	29
4.3.1. Discussion on results of the online tools scenarios	30
4.4. The online tools scenarios vs. the local scenarios.....	30
4.4.1. Discussion on online tools scenario vs. local scenario	31
5. Conclusion	32
5.1. Research overview	32
5.2. Limitation	34
5.3. Future work and recommendations:.....	35
5.4. Contributions.....	36
6. References.....	37

Acknowledgements

I would like to express my deep gratitude to my supervisor, Dr. Simon Scheider, for his continuous and full support during this graduation project. Special thanks to my responsible Prof. dr. Stan Geertman for the useful comments and kind support during my master's course. Next, I would also deeply thanks to the Geoforum community, in particular Dr. Wouter Beek and Dr. Erwin Folmer, who supported me with answering my questions about the PDOK infrastructures. I would also like to thank my family for their infinite love and support. Finally, many thanks to Dr. Ehasan Hamzei for his feedback.

Abstract

The volume of geodata available on Spatial Data infrastructures (SDIs) continues to grow, and there is an increasing problem with the abundance of geodata in terms of the discovery and accessibility in the distributed environments. It is difficult for end-users to find relevant content provided by different data providers. This problem becomes more challenging when it comes to supporting natural language in search engines; since the effectiveness and the findability of datasets rely on search techniques and the clarity of the queries. The keywords used by users are often different from the keywords recorded on metadata. However, the keywords submitted to the search engines may have semantically related to the content of metadata. Therefore, Natural Language Processing (NLP) techniques can be employed in conjunction with the technology used in the search engines to help different users with language limitations and specific domains by capturing the semantic and linguistic content in metadata. When a query executes poorly, the business logic behind the search engine reformulates and enriches queries based on the synonyms and relations gathered from the online data resources, which affects the recall and precision of geodata retrieval. This approach is a common technique and has been implemented for open-domain search engines such as Google and location-based services. However, spatial search and NLP techniques on the current Catalogue Services (CSs) are ongoing research topics and still required much work to be beneficial for users to take advantage of existing open government datasets. To address the limitations of search on the current SDIs and bridge between users' minds and contents documented in metadata, in this research, we examined query expansion using WordNet and Google translate API to generate more semantic keywords. In this work, we proposed a corpus-based methodology for query keyword extraction. The corpus is gathered from real users' questions in natural language. Then, these keywords are enriched using WordNet and Google translate APIs. Evaluation is carried out compared to a manual gold standard and baseline for the set of query keywords. Our empirical evaluation study on these tasks is executed by developing three retrieval platforms. Our study shows that the semantic keywords resulted from the local multilingual WordNet platform help users by reformulating good alternative queries. This approach also causes improvement in the precision and recall of geo-datasets by 1% and 22% respectively.

1. Introduction

1.1. Background

The PDOK and National geo-registry are national platforms and geospatial information brokers that facilitate the users' access to a broad collection of geographical datasets. The major challenge in current Spatial Data Infrastructures (SDIs) is scaling Natural Language techniques to a large number of distributed datasets to facilitate geodata retrieval. These techniques allow search engines to capture the semantic and linguistic content and assist users with language and domain limitations. Current SDIs employ metadata to describe, manage, discover, and exchange data and facilitate the discovery of spatial datasets based on users' queries (Chen et al., 2018). The geographic metadata provides descriptive information that can reduce the discovery of datasets in Spatial Data Infrastructures (SDIs). This descriptive information consists of the title, keywords, abstract, spatial domain, and time periods known as the attributes of geographic data (Chen et al., 2018; Chen & Yang, 2020). These attributes encompass geographic information in terms of text that can significantly improve the quantity of geodata source retrieval for queries in natural languages (Tóth, 2012).

On the other hand, the technologies used in search engines are language sensitive. It gets even worse when the keyword used in the query is semantically or linguistically different from the ones used in the metadata. More precisely, data providers are co-located and adjacent governments that describe datasets differently. This description information is often different from what is searched by the public consumers (Lafia et al., 2018). This problem is revealed in lower precisions and recalls of search results. However, in an ideal SDI, search engines are expected to cross domains and help users by capturing the semantic and linguistic content of datasets and metadata (Tan et al., 2006; Tóth, 2012, Gong et al., 2005).

The current search functions used in SDIs are exact-match keywords that refer to a perfect match between the keywords input by the users and the content of the geospatial resources. The exact-match keywords search method cannot deal with the ambiguity of natural language and semantic heterogeneity in user keywords. As a result, a new trend in research is a transition from keyword-based to semantic search. Semantic search aims to add semantic keywords by machine and match user queries to content in metadata. Thus, search engines can discover relevant datasets even though they are not labeled with the exact keywords in the metadata (Li et al., 2016; Lutz & Klien, 2006).

In this approach, the relationship between various kinds of geodata and the semantic relation of keywords and metadata contents are required to be understood. This understanding promotes semantic trackability in terms of lexical and structural similarity and has a major role in disambiguation (Scheider et al., 2020; Unger et al., 2014). To be more specific, a certain ambiguity in a geo-analytic question can be translated into geography phenomena. For instance, in the given question *"What is the average distance to green areas per PC4 area in Amsterdam?"*, a simple term, like a green area, can be referred to an object (e.g., greenhouse or park datasets) or a collection of objects (e.g., trees dataset). Therefore, the derivability of answers that are explicit or implicit (need additional reasoning) is an issue (Scheider et al., 2020; Unger et al., 2014).

The above-mentioned issues have attracted remarkable interest within Geographical Information Science (GIScience), in particular, geo-information retrieval, geoparsing, and natural language processing (Ballatore et al., 2013). There is an impressive number of semantic similarity and relatedness methods represented in various applications and domains (AlMousa et al., 2021). These methods have been tailored to measure and compute similarity and relatedness between concepts in different domains (Ezzikouri et al., 2019).

The most common solution for measuring similarity and relatedness between two concepts is using the ontology for a set of keywords and metadata to facilitate the discovery of datasets (Chen & Yang, 2020; Espinoza & Mena, 2007; Lutz & Klien, 2006). However, there are several problems to develop an ontology for the keywords embedded in questions and metadata. The first problem is ontologies may only leverage the discovery of limited geographic metadata (Chen & Yang, 2020). Another reason is that the ontology approach is a complex process in terms of geographic phenomena details and annotating datasets. Consequently, building an ontology that covers all terms and concepts is a time-consuming process and requires agreement between experts to define concepts and terms. Furthermore, ontologies are established based on existing documents. If documents are not complete and standard, this can hurt the hierarchic and thematic links. As a result, more common spatial language requires formalizing the knowledge (Billen et al., 2011).

To address cross-lingual information retrieval (CIR), the common approach is using translation APIs, such as Yandex Translate API, Google translate API, and Microsoft Text Translation API. These APIs allow users from different language backgrounds to find data and information from data repositories recorded in languages other than their native language.

Considering the aforementioned issues with conventional methods of geo-information retrieval in SDIs, this research aims to find techniques that not only are ontology independent but also improve users' queries by expanding keywords using cross-lingual and semantic NLP techniques. This research is organized into 5 chapters. Chapter 2 reviews the literature. Chapter 3 focuses on the methodology and implementation, chapter 4 is the results and discussion. The last chapter is the conclusion.

1.2. Objects and research questions

This research sets out to investigate the potential of natural language processing (NLP) techniques to enhance the quality of geodata source retrieval in SDIs using semantic keywords for the geographic phenomena requested. Geodata source retrieval addresses the problem of finding those datasets whose contents match a user's request directly over metadata in spatial data infrastructure. Queries are selected from a large geographic question dataset (geo-analytical question corpus) in which keywords are extracted that denote different geographic phenomena. Query keywords are then expanded using Google Translate, and WordNet to capture the semantic context of the keyword. Evaluation is done with respect to a manual gold standard, a baseline, and information retrieval (IR) metrics for the set of questions and the metadata retrieval. Then, the quality of the keyword-based approach is compared for defined scenarios. In the future, this helps us to a step towards using data and geo-computational resources for recommending suitable data and analysis for geo-specialists (Scheider et al., 2020). Therefore, the main research question is:

"What types of NLP techniques can be used to improve the findability of geodata sources, and to what extent are NLP techniques efficient? "

To answer this question, the following sub-questions are to be covered:

- *How can NLP techniques be used to retrieve and search over metadata?*
- *To what extent can multi-linguistics problems be handled using Google API?*
- *To what extent WordNet is effective for the query expansion method?*
- *Which query expansion method is more suitable for the proposed corpus?*
- *How much does the result of keyword expansion promote retrieval quality?*
- *To what extent the proposed corpus and the NLP techniques are effective and efficient for the online services on the infrastructures?*

1.3. Scope of research and limitations

This section explores both the scope of the research and the limitations in each step of the methodology. Although this thesis is grounded on the research progress that has been conducted by Wieleman, 2019; Xu et al., 2020, it is not about how to collect and analyses question corpora, nor the semantic structure of spatial questions. To summarize, in this research, we are looking for semantic keyword expansion methods based on geographic phenomena that can promote catalogue services to be more user-oriented in the future and retrieve more relevant metadata in the geography domain. Although using ontologies can be beneficial in terms of enhancing relevant geodata retrieval, it is beyond the scope of the current thesis. This research does not explore the development of an interface for capturing spatial questions and dynamic map content creation or spatial analysis since designing and implementing a user-friendly interface can be a separate thesis. Figure 1.1 demonstrates the general architecture of a retrieval platform in a real retrieval system that may be part of a search engine. In this thesis, we only focus on the retrieval platform. The retrieval platform consists of two separate components. The question processing is carried out manually, and in the future, keyword extraction will be automatic since it is defined as a separate project. Therefore, when users ask a question in a natural language (e.g., *What is the average distance to green areas per PC4 area in Amsterdam?*), algorithm(s) analyzes the question and extracts keywords automatically from the question. The second phase is query processing consist of an algorithm(s) that allows query expansion and reformulation, and the last phase is geo-data retrieval.

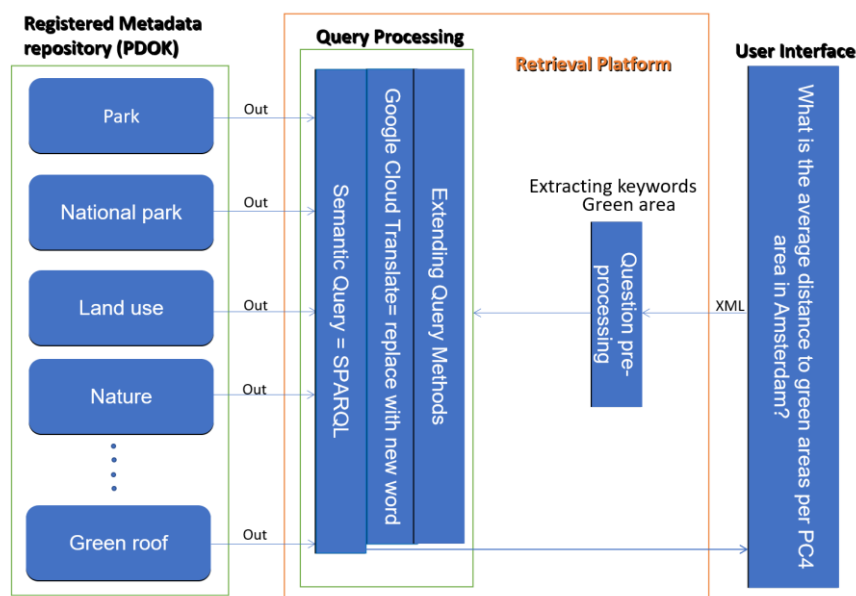


Figure 1.1: High-level components of a retrieval platform

2. literature review

Over two decades, a lot of research has focused on information retrieval and researchers have devoted their studies to use natural language processing. They have introduced various methods for question answering using linguistic, statistics, and programming techniques (Bucher et al., 2020; Jones & Purves, 2008). On the other hands, large online data sources (e.g., WordNet, Frame Net, Wiktionary, YAGO2, and DBpedia) facilitate query expansion (Bao et al., 2014; Cai et al., 2020; Chen & Yang, 2020; Leseva et al., 2018). Essential steps for information retrieval are: (1) the analysis of questions into terms or keywords from geo-analytic question corpora, (2) the mapping of phrases (e.g., entities) to the Knowledge-Base(KB) or Knowledge Graph(KG), (3) entity disambiguation, and (4) query construction and (5) query execution over the KB or KG (Scheider et al., 2020; Diefenbach et al., 2018). This chapter explores existing approaches and provides backgrounds for the proposed methodology.

2.1. Geo-analytic question corpora and analyzing geographic questions

A key requirement of retrieval systems is the availability of corpora that may drive from different sources, such as papers, textbooks, online content, and more. Geo-analytic questions are mostly collected from corpora and provide knowledge about spatial patterns and relations that often do not have an obvious answer. These questions are ranked from simple questions (e.g., a question about location) to complex questions (e.g., a question about the condition, pattern, trend modelling, and what-if modelling) (Xu et al., 2020). The answering process starts with a disintegrating question into multiple simple parts or keywords that may be answered separately.

The first step to extract keywords is acquiring semantic information from questions. The semantic information types may be a place name, spatial relationship, place types, and more others. In information retrieval, discriminating between the entity intent is imperative to parse the questions. From a geospatial viewpoint, *"most frequent entity intents are pattern, relationship, distribution, density, effect"* (Xu et al., 2020). For example, given the question *"What is the concentration pattern of ethnic groups in Amsterdam in 2019?"*, "the concentration pattern of the ethnic group" is the intent phrase restricted with a place name "Amsterdam" and time "2019". In some cases, the intent word in a question may be accompanied by one and more adjectives, adjectival nouns, or noun modifiers. For a given question, "What is the population density in Randstad? ", the intended word is "density" accompanied by the attributive nouns "population". In geo-analytic question answering, the intent of questions can be considered an answer. Moreover, other words such as population, land use, and ethnic that refer to a set of geographic phenomena or objects (spatial and non-spatial datasets) are relevant to answer (Xu et al., 2020; Scheider et al., 2020). This set of geographic phenomena are content requirements for answering the question.

2.2. Query expansion

Query expansion is known as a process of selecting and adding terms to the user's query to reduce query-document mismatches (Flank, 1998; Buscaldi et al., 2006; Vechtomova et al., 2003). Query expansion methods allow that the original query is reformulated. These methods help search engines to find synonyms of words, mapping and re-weighting the terms, measuring semantic similarity, and relatedness. More precisely, algorithms help terms to be extracted automatically from knowledge resources (e.g., thesauri) or documents. This process allows the algorithm to find a stronger semantic association with the original query and discriminate between the relevant and irrelevant documents (Chen & Yang, 2020).

Consequently, the search engine can cope with the mismatch problem and increases retrieval performance by improving a short and incomplete query (Pivert & Smits, 2020).

Several techniques and methods have been proposed for query expansion. These methods mostly employ two or more combinations of statistics, linguistics/semantics techniques, and artificial intelligence or heuristic algorithms (Mitra et al., 2019; Mai et al., 2019; Chen & Yang, 2020; Elbedweihy et al., 2013). In this section, we only focus on linguistics techniques developed using WordNet and Google translate.

2.2.1. Query expansion using WordNet

WordNet is a popular semantic data resource used for many applications in NLP and computational linguistics since 1990 (Ballatore et al., 2013). This online lexical data source resembles a thesaurus used for anchoring different types of semantic knowledge for concepts by grouping them into sets of cognitive synonyms called "synsets". It also covers the four-common part-of-speech tags (or POS tags) consists of a verb, noun, adjective, and adverb (Laparra et al., 2010; Perkins, 2014; Lu et al., 2015). This data source allows concepts to be organized into a conceptual hierarchy by interlinked synsets using conceptual-semantic and lexical relations, in which they are categorized based on synonyms and taxonomic relationships (AlMousa et al., 2021; Lu et al., 2015). Keyword expansion can be computed along each dimension using algorithms and specific senses of words using similarity distance between concepts (Gong et al., 2005; Laparra & Rigau, 2009; Elbedweihy et al., 2013; Ballatore et al., 2013). The similarity task is used as an intermediate task for query relaxation in geo-data retrieval (Ballatore et al., 2015).

More precisely, in WordNet, semantic distance computes the similarity between two concepts with the same lemma (a form of the word, e.g., two nouns) based on the shortest path between concepts. The similarity measurement uses the number of nodes to compute the semantic distance between concepts (Pedersen et al., 2004). The result of a semantic similarity measurement is quantified and represents a real number normalized in the interval between 0 and 1, known as a similarity score. The similarity scores are meaningful and provide useful information when a concept is compared to other concepts (Ballatore et al., 2015). The concepts are compared based on the defined relationships in WordNet. These relationships consist of 1) synonyms (same and interchangeable concepts) 2) hyponymy (sub-type), and hypernymy (super-type) 3) meronymy and homonymy (part-whole), toponymy (which indicates manners), and antonymy (opposites concepts) (Degbelo & Teka, 2019).

There are four approaches for query expansion in WordNet. A common approach in information retrieval for query expansion is replacing the keyword in the original query with its set of synsets (Gong et al., 2005; Lu et al., 2015; Degbelo & Teka, 2019). This method was examined by Lu et al., 2015 and enhanced the precision and recall of relevant documents on 20 search tasks, by 5% and 8%, respectively. The second and third approaches are similarity and relatedness computation between concepts that more information is explained in sub-sections. The last approach is hybrid methodology. In this work, we examine the combination of these approaches and evaluate the effectiveness of our approach for our case study.

2.2.1.1. Similarity methods

There is an impressive number of methods to measure similarity and relatedness between two concepts in WordNet (cf. Table 2.1). To select the more suitable method, the similarity measurement is usually evaluated based on concept similarity ratings by human judgment and

if a measure mimics human judgment, then the similarity measurement can be used for query expansion in geographic information retrieval. Ballatore et al., 2015 calculated the correlation between the 97 pairs of OSM concepts with human judgment and concluded the existing WordNet-based similarity measures are not sufficient to compute the semantic similarity in their case study. Their experimental results showed that hso, vector, and vectorp ($\rho = [0.43, 0.53]$) similarity methods are the top-performing measurements with higher cognitive plausibility. The second place was for the path similarity family, and the other measurements achieved a lower performance ($\rho < 0.34$) (Ballatore et al., 2015). Another approach is task-based evaluation for each path similarity method to quantify the effectiveness of queries using known IR metrics (Ballatore, 2013). The approach has been elaborated on section 2.3.

Name	Description	Name	Description
path	Edge count	wup	Edge count between <i>lcs</i> and terms
lch	Edge count scaled by depth	hso	Paths in lexical chains
res	Information content of <i>lcs</i>	lesk	Extended gloss overlap
jcn	Information content of <i>lcs</i> and terms	vector	Second order co-occurrence vectors
lin	Ratio of information content of <i>lcs</i> and terms	vectorp	Pairwise second order co-occurrence vectors

Table 2.1: WordNet-based similarity measures (Ballatore et al., 2015).

2.2.1.2. Relatedness between concepts

Although semantic similarity and relatedness are very closely related, semantic similarity should not be confused with semantic relatedness. Indeed, semantic similarity is known as a specific type of semantic relatedness in an is-a relation (Ballatore et al., 2015). For example, "school" is semantically related to "building", while "school" and "university" are semantically related and similar. Therefore, "school" and "university" may represent a higher similarity (0.57) than the similarity between "school" and "building" (0.13).

Relatedness is mostly a heuristic methodology used or designed by different researchers. One approach is computing the relatedness between the synsets and glosses (a brief definition) of the two concepts based on the set theory concepts. This approach has been introduced as a novel and optimal approach by Ezzikouri et al., 2019 to improve the search of relevant information for each domain. In this method, each term's synsets and its equivalent gloss are computed using WordNet. Then, the intersection and union between the pairs of synsets and gloss are computed to find the score of the similarity. This method has been proposed without representing any evaluation and empirical results. The relatedness approach is also a common approach to address word sense disambiguation (Aouicha et al., 2018). Aouicha et al., 2018 proposed extra steps in the methodology by incorporating gloss of synsets in WordNet and Wiktionary. The results of gloss were expanded in different dimensions. The experimental results are compared with the results of the methodology using only WordNet and represent an improvement.

2.2.2. Cross-lingual information retrieval

Cross-lingual information retrieval (CLIR) systems enable users to search and find their required data and information from data repositories recorded in languages other than the user's native language. As a result, users can overcome the language barrier problem. Google translate benefits from statistical analyses that provide better results to translate hierarchically structured terms (e.g., ontology). This API can distinguish labels, detect language, and translate them directly (Florence, 2020; Lin & Krizhanovsky, 2011). Google translate API has been used by a lot of researchers to address the multilingual issue with customization of plugins for visualizing ontologies (Florence, 2020), translating biomedical ontologies

(Bouscarrat et al., 2020), ontology matching (Lin & Krizhanovsky, 2011), information retrieval for the medical domain (Rahmani et al., 2017), information retrieval (Segev & Gal, 2008), and creating translation chain (Sequeira et al., 2020). Then, the results of the experiment were examined with other available tools such as the Wiktionary database. The accuracy of google API was above 60 percent (Lin & Krizhanovsky, 2011). Furthermore, the result of the translation chain into various languages showed that near languages in the same family (e.g., Dutch, English) preserve high accuracy above 86 percent (Sequeira et al., 2020).

2.3. Evaluation methods

Evaluation is a vital and tedious task in the context of information retrieval. The evaluation process is carried out against a source of data to study the strengths and weaknesses of search engines and QA systems using a gold standard. More precisely, gold standards are an essential resource for assessing NLP systems (Deleger et al., 2014, Zuva & Zuva, 2012). With a gold standard, researchers aim to evaluate and measure the quality of the linked data, ontologies, question answering systems, or platforms. In IR, a gold standard is a set of correct answers to a query (Fathalla et al., 2019; Brewster et al., 2004; Deleger et al., 2014; Usbeck et al., 2018; Sun et al., 2019). Furthermore, this evaluation system facilitates reusing and developing new models or platforms identified by other computational experts (Costa et al., 2020; Fathalla et al., 2019; Brewster et al., 2004; Deleger et al., 2014).

In literature, many retrieval models, algorithms, and systems are assessed to improve systems. What important is to compare the relevancy of queries to items. Different algorithms and indices are used to evaluate retrieval platforms (Zuva & Zuva, 2012). In the area of natural language processing and information retrieval, task-based evaluation is applied to study specific tasks over question answering platforms. The task-based evaluation is used to quantify the effectiveness of a task considering known IR metrics (Ballatore, 2013). These evaluation metrics are based on the Cranfield paradigm for information retrieval in document collection and consist of recall, precision, and F-measurement for unranked retrieval sets (Ceri et al., 2013, Zuva & Zuva, 2012).

The recall represents the ability of a retrieval platform in finding relevant documents, whereas precision demonstrates how good a platform is to retrieve only relevant documents (Mandl, 2008). F-measure is defined as the harmonic mean of precision and recall. F-measure assesses precision/recall trade-off (Sasaki & Fellow, 2007).

The mentioned IR metrics are calculated with the formula 1, 2, 3:

$$Recall = \frac{RRE}{ARE} \quad 1$$

$$Precision = \frac{RRE}{RE} \quad 2$$

Where ARE is the total number of relevant answers for keyword A in the gold standard. RRE is the total relevant documents for keyword A. RE is the total retrieved documents (relevant documents \cup irrelevant documents).

$$F - measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad 3$$

3. Methodology and implementation

3.1. An overview of the methodology

The proposed methodology consists of six main phases. Figure 3.1 shows the pipeline of the proposed methodology. The first phase is keyword gathering. This phase is mainly carried out manually to extract query keywords from corpus questions. The second phase is the answer dataset in which metadata set and metadata keyword set are gathered. Metadata set is employed for the scenarios, and metadata keyword set is used for WordNet scenario in the fourth phase. The gold standard is the third phase in which a manual goal standard is established.

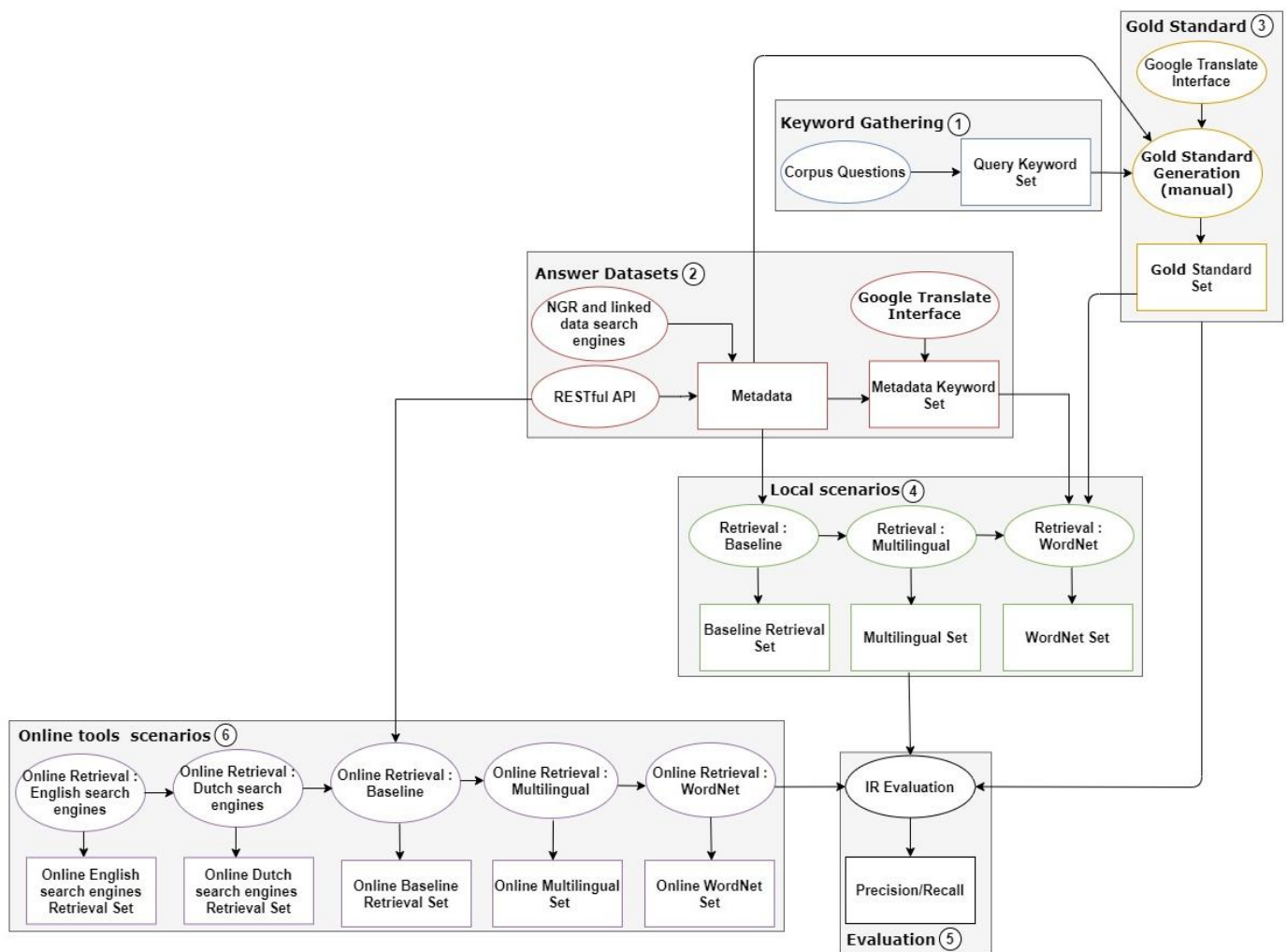


Figure 3.1: The proposed methodology

The fourth phase represents three scenarios and is named local scenarios. These scenarios are retrieval baseline, retrieval multilingual, and retrieval multilingual WordNet scenarios. The fifth phase is the evaluation step which IR metrics are defined to assess the retrieval results. Lastly, five online scenarios are defined to assess the online services and compare the results with the local scenarios. These scenarios consist of online retrieval baseline, online retrieval multilingual, online retrieval WordNet, online Dutch search engines, and English

search engines. All relevant code and documents can be found at [https://github.com/mariamsajadian /query-expansion_thesis](https://github.com/mariamsajadian/query-expansion_thesis).

3.2. Dataset and infrastructures

This section explores two primary infrastructures that are required for the overall methodology. We represent the proposed dataset for keyword gathering and select a platform for metadata gathering. In this work, we propose a dataset that consists of geo-analytic questions in an excel file. Xu et al., 2020 have introduced a new corpus named GeoAnQu. The selected dataset consists of a real question dataset from real users. This dataset consists of approximately 429 geo-analytic questions extracted from 100 scientific papers and English textbooks. These articles were sieved by three criteria: 1) the field of Human Geography, 2) containing GIS analysis, and 3) published in 2009-2018. The second source of dataset is textbooks on GIScience and GIS: David, 2010; Heywood, 2011; Kraak & Ormeling, 2013; O’Looney, 2000.

In this thesis, two infrastructures have been considered to harvest machine-readable datasets in RDF and XML formats. These platforms are two national brokers called PDOK and NGR. These platforms play a role as a third party between data and service providers with end-users. Metadata that facilitate users' access to a broad collection of geographical datasets has been distributed in heterogeneous formats: XML and linked data. These platforms embrace different search engines: two classic search engines, an Elastic search engine (Kadaster, 2020.a), a browser search engine (Kadaster, 2020.b), a keyword search engine (Kadaster, 2020.c), and a search engine for linked data (Kadaster, 2020.d). Recently, a SPARQL endpoint has been developed that allows the accessibility of end-users through RESTful API to metadata. The responsible service provider is the Kadaster agency that ensures the findability of data via a SPARQL endpoint.

In this work, three services are employed for data gathering. These services are RESTful API, the linked data search engine, and the classic search engine on NGR. RESTful API is used for keyword extraction from metadata. The linked data search engine on PDOK and the classic search engine on NGR are, respectively, used to gather metadata in the linked data (11405 triples) and XML formats. Lastly, the metadata is stored on the local machine used for the local scenarios and the evaluation phases. Moreover, RESTful API is used for the online scenarios described in the sixth phase.

The selection of services is made after comparing services and datasets. This comparison was conducted considering accessibility to new services and up-to-date metadata. The PDOK platform provides an impressive number of services that are compatible with new technologies, such as RESTful API and linked data. Linked data is an innovative approach that attracted the attention of many scholars and is predicted to become a popular industrial technology. Furthermore, linked data is a great technological solution for today’s organizational problems and publishing data sources on the web (Folmer et al., 2020). Moreover, RESTful APIs are widely used in the modern web; since they are efficient for various applications (Khodadadi et al., 2015). We also used the NGR search engine to enrich metadata to cover more keywords. The preliminary experimental evaluation of search engines shows us that the total number of retrieved metadata using RDF metadata is lower than the NGR search engine. As a result, metadata was enriched with extra metadata gathered in the XML format, and the total number of triples increased from 11405 to 11914.

3.3. The first phase: Keyword gathering

The GeoAnQu dataset consists of different questions that might be asked by various experts. The questions can be the subject of different study areas in geography (i.e., hazard management, tourism management, water management, health and liveability, geographic criminology, climatology, and more others). Table 1 shows a sample of these questions and keywords. The questions are linguistically complex and are required to be customized to search engines. Users usually perform search queries by adopting questions to keywords. These keywords contain geographic phenomena that are relevant to the answer. More precisely, geographic phenomena show the form and the contents of the questions. The contents are extracted as the keywords to be part of the answer to the geo-analytics questions. Additionally, these keywords may represent similarity and relatedness with metadata attributes. Also, these keywords provide the search engine with general knowledge and make the subject revolving around geographic phenomena in a period of time or/and place.

Question	Keywords
How many people are affected by a hurricane in Oleander?	People – hurricane
What is the average distance to green areas per PC4 area in Amsterdam?	Green areas
Which park does have the highest concentration of the bald eagle in Texas?	Park – eagle
How did tornado strength and wind speed change over time on April 2011?	Tornado – Wind

Table 3.1: Sample of GeoAnQu dataset

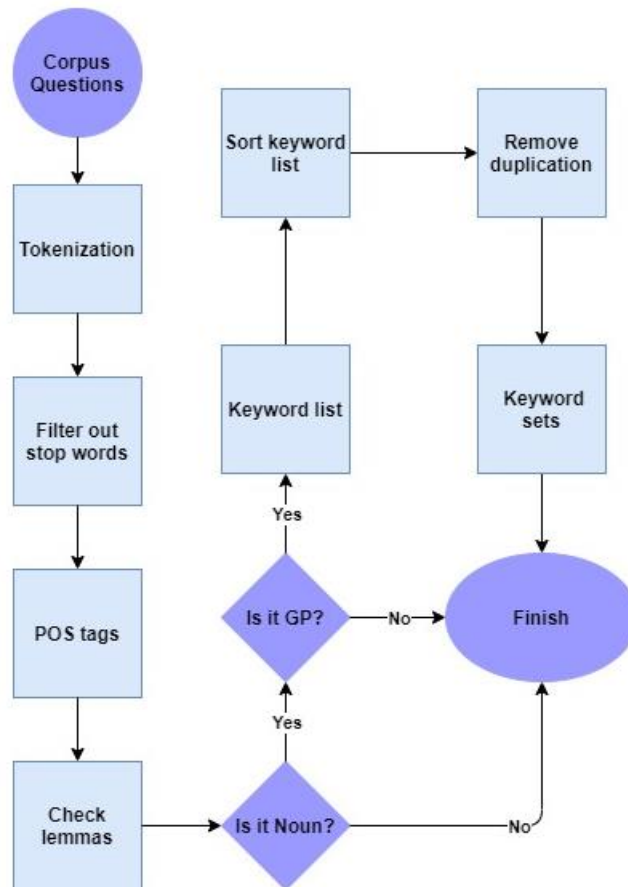


Figure 3.2: Process of keyword extraction

In this research, although we did not use WordNet to extract keywords, we mimicked the process of keyword extraction in a search engine. Figure 3.2 represents the process of keyword extraction. To start with, tokenization is a task to split sentences into individual terms or words, called tokens. The next step is removing stop words or function words (have little lexical meaning) and may cause a problem in a search engine, such as the existence of prepositions, articles, auxiliary verbs, WH question words, and more. A POS tag is a grammatical tagging and indicates the part of speech to discriminate between nouns, verbs, adjectives, and adverbs. In this research, keywords, and synonyms with the same POS (nouns) are returned using the lemma (a form of keywords). Nouns refer to a person, places, things, or concepts, such as geographic phenomena (GP) that show the distribution in space. In this research, place names were excluded, and the focus is only on geographic phenomena. To create the final keyword list, different criteria have been considered. These criteria consist of the frequency of occurrence and distribution of keywords for the set of selected questions. For example, "house", "road", "hurricane" keywords have been repeated in 10, 15, 9 different questions, respectively. The general frequency of occurrence of a word represents how important the keyword is to a document in a corpus. However, in the final dataset, the duplicate keywords have been excluded from the list to avoid redundancy. The output of this step is the keyword sets in English. The keyword sets represent various geographic phenomena that refer to man-made and natural phenomena and the subject of different study areas in geography (i.e., natural catastrophes, urban infrastructures, tourism management, water management, health and liveability, geographic criminology, climatology, biogeography, and more others).

3.4. The second phase: Answer dataset

As mentioned in section 3.2, the linked data search engine on PDOK and the classic search engine on NGR are, respectively, used to gather metadata in the linked data (11405 triples) and XML formats. The result of metadata gathering is an enriched RDF file that contains 11914 triples. Besides, RESTful API and Python codes are used for keyword extraction from metadata. These keywords are used as a dataset to measure similarity and compute the semantic overlay in the WordNet scenarios (cf. Section 3.6.3). Each extracted keyword from metadata is manually translated into English using the Google translate interface and documented in an excel file. The outputs of this phase are RDF metadata (11914 triples) and metadata keyword sets (252 keywords).

3.5. The third phase: Gold standard

In this section, we report on establishing the manual gold standard. The gold standard is a test document and contains query keywords and the numbers of the relevant answers for each query keyword. By the gold standard, we aim to evaluate the quality of retrieval scenarios in terms of recall and the precision of the translation system. The prerequisites of the gold standard are the outputs of query keyword sets and the RDF dataset explained in sections 3.3 and 3.4. In this step, the query keywords are translated into Dutch using the Google translate interface (cf. Figure 3.1). Then, different synonyms of keywords were manually searched in the RDF file to find the best match with metadata on the local machine. Additionally, the keyword search engine and the NGR search engine are used to validate the results of the gold standard. The result of this phase is a gold standard that covered 167 keywords in English and Dutch and the total numbers of relevant answers for each query keyword in the RDF dataset.

3.6. The fourth phase: Defined local scenarios

This section aims to define three scenarios and examine query keywords reformulation over metadata. The first scenario is the baseline and known as a benchmark for the evaluation and a query platform for other scenarios. The second is a multilingual scenario built on the top of the baseline to generate an automatic translation system. Lastly, the multilingual WordNet scenario is considered a mature search package and consists of the baseline and multilingual scenarios. The following subsections elaborate on each of these scenarios.

3.6.1. Baseline scenario

A retrieval baseline platform can be a naïve or a smart (strong) system used as a benchmark for comparison (Dalianis, 2018). In this work, the baseline is a naïve retrieval platform developed to examine the query keyword list over metadata. The scenario is to study more faithful queries to what users intend and to find the original query keywords without any query manipulation by the machine. Moreover, the baseline scenario is the building block for other scenarios. Figure 3.3 represents the main components of the baseline platform. The baseline consists of Dutch query keywords, RDF metadata on the local machine, Python codes, and baseline dataset. The Dutch query keywords in the gold standard are used to query over metadata. More precisely, queries are executed in the SPARQL language; and simple text matching is used against RDF datasets to retrieve datasets using Python codes. The output of this step is the baseline retrieval set. In this file, there are two classifiers for each keyword that consists of the number of relevant and irrelevant retrieval results. The results of the baseline are used to assess the query expansion scenarios (cf. Section 3.8).

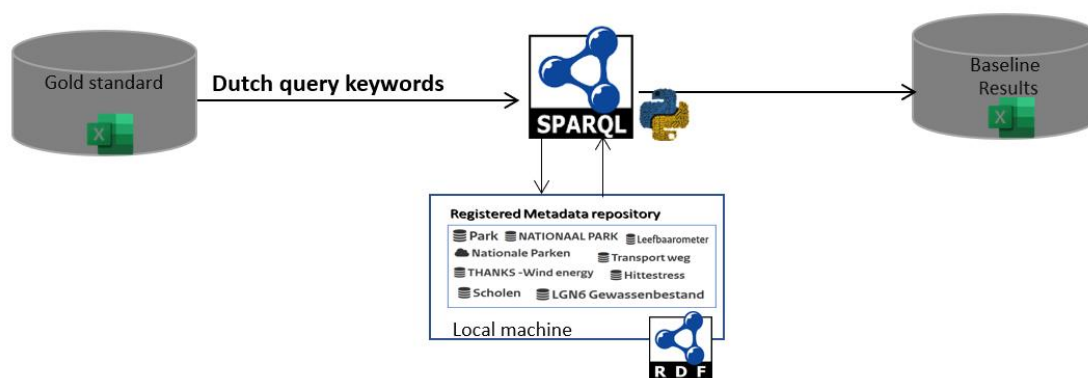


Figure 3.3: Components of the baseline platform

3.6.2. Multilingual scenario

The second scenario provides a base to cross the monolingual baseline to the multilingual platform using Google translate API. In the gold standard, the Dutch query keyword is gathered using the Google translate interface. These semantic keywords were searched manually based on the trial-and-error method over the local metadata to establish the gold standard. In this scenario, we replicate the experiment automatically using Python codes and the Google translate API that allows the query keywords to be translated and matched with metadata attributes on the local machine. Figure 3.4 demonstrates the components of the multilingual platform. This scenario is built on top of the baseline; additional codes are developed to facilitate accessibility to the translation API. The English query keywords recorded in the gold standard are queried over the multilingual system. The results of this task are compared with the results of the gold standard and the baseline to measure the precision

of the translation and retrieval results, respectively. The output of the multilingual scenario is an excel file named Google translate API retrieval set. In this file, there are two classifiers (the total numbers of relevant and irrelevant retrieval results) for each keyword and translation results. Please note that although Google translate API allows language detection and keyword translation, the translation API (unlike the Google translate interface) does not provide different synonyms for each keyword.

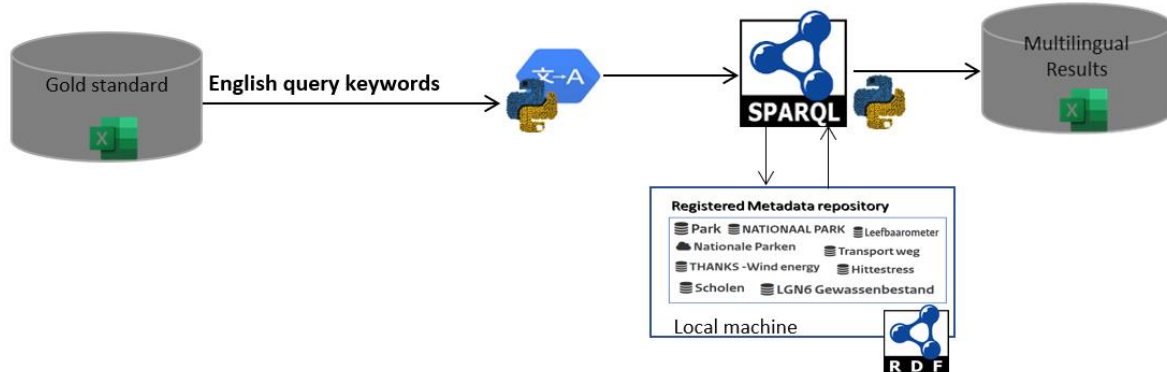


Figure 3.4: Components of the multilingual platform

3.6.3. WordNet scenario

This step is the last approach to address mismatch queries and prevent users from being faced with an uninformative result. As shown in Figure 3.5, the baseline and the multilingual platforms are reused to examine the recall of query expansion in the English WordNet. Query expansion in WordNet is carried out in five phases. The first phase represents hierarchal relations and computing synsets (i.e., synonyms, hypernyms, and hyponyms) of keywords. The second and third phases are computing the similarity and semantic overlay, respectively. Next, the query expansion results are translated into Dutch. Finally, the SPARQL query is executed against RDF metadata, and the query results are recorded in the WordNet scenario retrieval set file. The file consists of two classifiers (the numbers of relevant and irrelevant retrieval results) for each keyword.

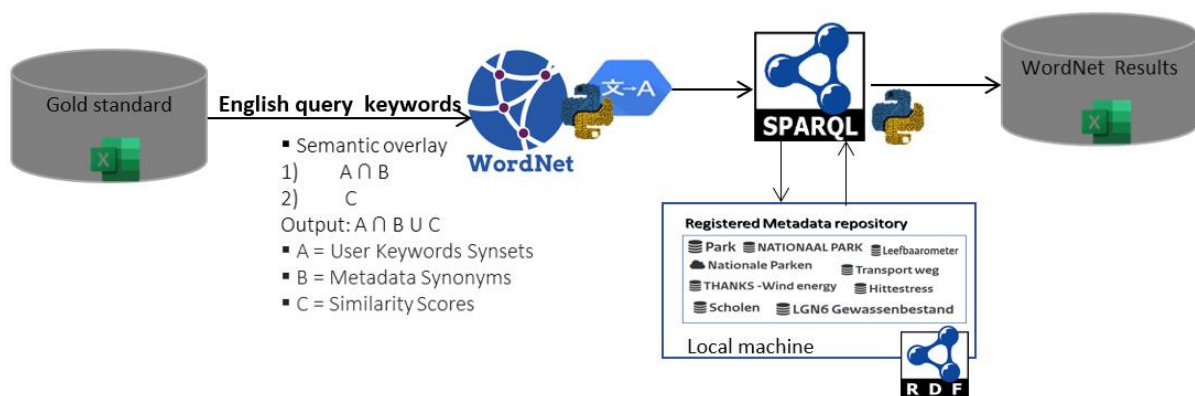


Figure 3.5: Components of the multilingual WordNet platform

3.6.3.1. Computing synsets

The keyword dataset represents semantic relations between query keywords in the gold standard with the metadata keywords explained in sections 3.3 and 3.4. Figure 3.6 shows the relationship between the query keywords and the metadata content. As can be seen, there is a semantic similarity relationship in an is-a relation between "eagle", "animal", "fauna", and

"frog", and "tornado", and "wind" as well as a semantic relatedness between "park" and "green", "people", and "population". In WordNet, these semantic relationships are seen in the three dimensions of synonyms, hyponyms, and hypernyms.

Figure 3.7 illustrates the hierarchal relation of "park", and "school" in the three dimensions. By replacing "park", "school" with "green" and, "primary school" respectively, the results of retrieval are enhanced. Therefore, the certain ambiguity for a simple term, like "green area" can be referred to "park" dataset. This pattern has been studied for all keywords, most of the matched query keywords with metadata are distributed in hypernym (super name) and hyponym (sub-name), and some cases are in synonyms. We also compute only the synonyms of metadata keyword sets to provide more semantic keywords and context for Google translate API. The semantic keywords context allows Google translate to deal with ambiguity. The synsets method is not useful for all query keywords. For example, "tornado" can not be matched with the "wind" dataset since wind is not defined in the hierarchal relation. To fill this gap, we use the similarity score explained in the next section.

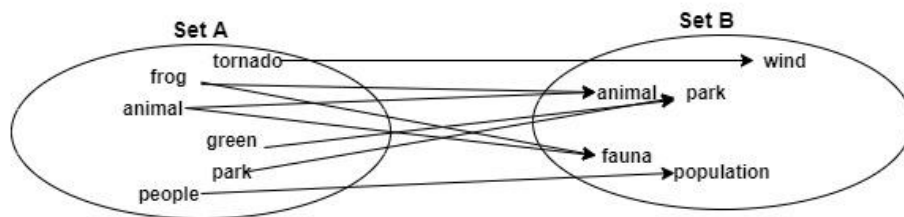
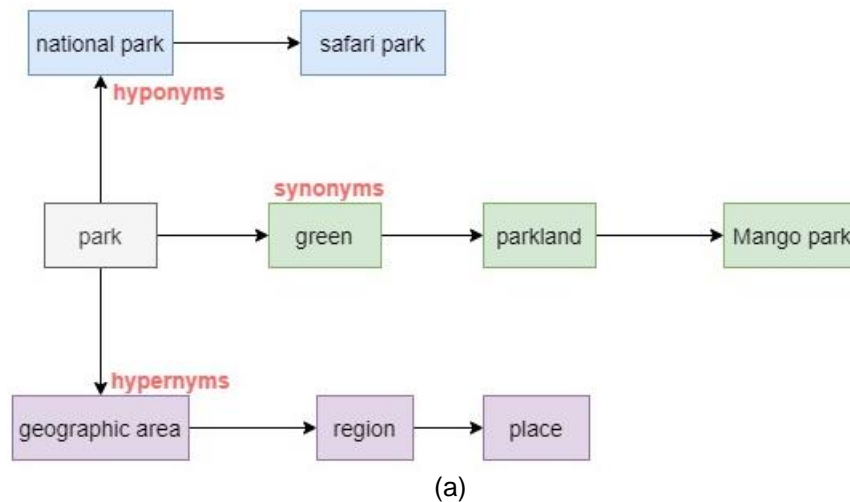


Figure 3.6: Relationship between query keyword (set A) and metadata (set B)



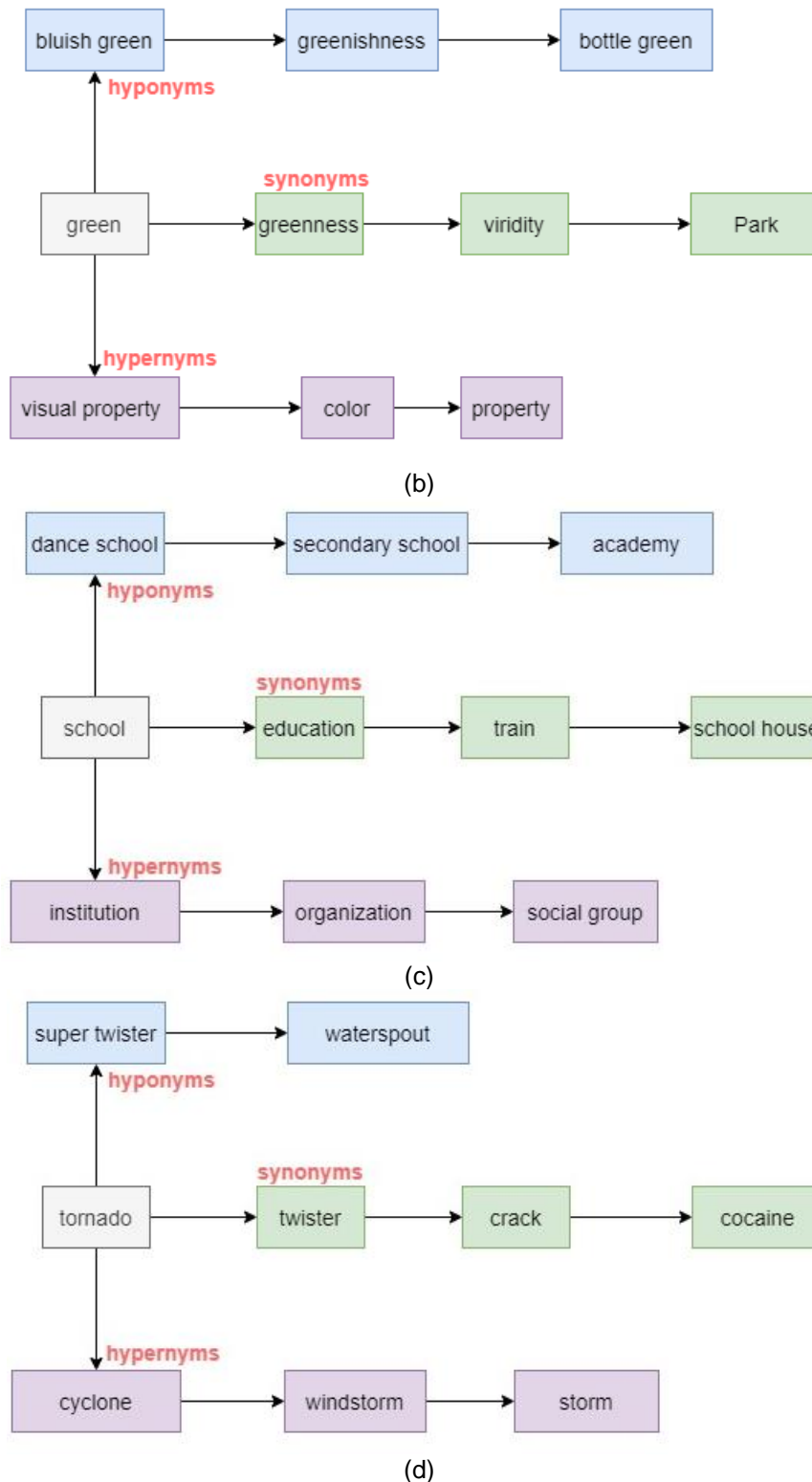


Figure 3.7: Hierarchical relation of "park", "school", and "tornado" in WordNet

3.6.3.2. Calculating similarity in WordNet

WordNet offers a different family of algorithms that can be used to measure the semantic similarity distance between two words, such as shortest path (path, lch, wup) and information content known as IC (res, lin, jcn). In this work, we only focus on path similarity algorithms implemented in WordNet. The path similarity methods are selected based on the results of

research by Ballatore, 2013. His work showed that after relatedness algorithms, path similarity algorithms are closer to human judgment. The relatedness methods are the hso (Hirst & St-Onge,1998), lesk (Banerjee & Pedersen, 2003), and vector (Patwardhan, 2003) implemented in Perl. Therefore, they have been excluded from our research. In this work, the path similarity is selected to capture the similarity between keywords.

To decide which path similarity measurement is the best match with our study, two steps are considered. In the first step, for each similarity method, a pairwise comparison matrix is devised to measure the similarity between keywords. Then, the results of the matrices are compared with all path similarity methods. Ich similarity shows an interval between [0, 3.637] that is normalized between [0, 1]. Table 3.2 represents the comparison between the path similarity family between query and metadata keywords. "School", "wind" , "plant", and "park" are metadata attributes, and "tornado", "high school", "green" are query keywords. The green color refers to the highest similarity between two keywords in each table, and the yellow color represents the irrelevant similarity score between two keywords. As can be seen, the similarity scores do not show the relatedness between "green" and "park", whereas this relation is presented in the hierarchical relation (cf. Figure 3.7 a, and b). Although path similarity algorithms may not distinguish between two related keywords (e.g., green, park, and plant), the wup and Ich methods represent a better similarity score compared to the path measurements for other query keywords.

Second, A task-based evaluation is performed for each similarity method. In this phase, similarity methods are examined to compute the similarity distance between 20 sets of pair geographic keywords, and the results of the computation are compared. Two criteria are considered to evaluate the task-based evaluation: the precision of the retrieval results, and the completion time (response time). The wup method represents slightly better results in the completion time compare to Ich. For example, the completion time to measure the similarity distance between two keywords by wup is 4 seconds. However, the same keywords take 11 seconds by the Ich method. As a result, the wup method is used to measure the similarity distance between two keywords. Moreover, the wup algorithm proposed by Wu and Palmer, 1994 counts edge and considers the longer path to the root node (ancestor node) where multiple candidates are available.

Path	school	high_school	wind	tornado	green	plant	park
school	1	0.33	0.0625	0.0555	0.0769	0.066	0.066
high_school	0.33	1	0.055	0.05	0.0666	0.0588	0.0588
wind	0.0625	0.055	1	0.1428	0.0625	0.071	0.0714
tornado	0.055	0.05	0.1428	1	0.05555	0.0625	0.0625
green	0.0769	0.0666	0.0625	0.05555	1	0.066	0.066
plant	0.066	0.058	0.071	0.0625	0.066	1	0.0909
park	0.066	0.0588	0.0714	0.0625	0.066	0.0909	1

(a)

wup	school	high_school	wind	tornado	green	plant	park
school	1	0.8888	0.1176	0.105	0.25	0.125	0.125
high_school	0.8888	1	0.105	0.095	0.222	0.111	0.11111
wind	0.1176	0.105	1	0.7	0.1052	0.235	0.235
tornado	0.105	0.095	0.7	1	0.105	0.2105	0.2105
green	0.25	0.222	0.1052	0.105	1	0.125	0.125
plant	0.125	0.1111	0.235	0.2105	0.125	1	0.375
park	0.125	0.11111	0.235	0.2105	0.125	0.375	1

(b)

lch	school	high_school	wind	tornado	green	plant	park
school	3.63	2.53	0.86	0.747	1.07	0.929	0.929
high_school	2.53	3.63	0.747	0.64	0.929	0.804	0.8
wind	0.86	0.747	3.63	1.69	0.86	0.998	0.998
tornado	0.747	0.64	1.69	3.63	0.747	0.86	0.86
green	1.07	0.929	0.86	0.747	3.63	0.929	0.929
plant	0.929	0.8	0.998	0.864	0.929	3.63	1.239
park	0.929	0.8	0.998	0.86	0.929	1.239	3.63

(c)

lch_norm	school	high_school	wind	tornado	green	plant	park
school	1	0.696	0.23	0.2	0.29	0.25	0.25
high_school	0.696	1	0.21	0.17	0.26	0.22	0.22
wind	0.23	0.21	1	0.46	0.24	0.27	0.27
tornado	0.2	0.17	0.46	1	0.21	0.24	0.24
green	0.29	0.26	0.24	0.21	1	0.26	0.26
plant	0.25	0.22	0.27	0.24	0.26	1	0.34
park	0.25	0.22	0.27	0.24	0.26	0.34	1

(d)

Table 3.2: Pairwise comparisons matrix

3.6.3.3. Calculating semantic overlay

After computing synsets of query keywords, the intersection between a set of synsets and the metadata synonyms is computed to maximize the number of common semantic keywords between the two keywords. Moreover, the intersection allows filtering out the semantic keywords that may be less relevant between the query and metadata and cause problems and noise for the translation system and retrieval results. Lastly, the results of the intersections are combined with the results of similarity to form the union. Figure 3.8 represents the results of two words, "hotel" and "building" in the Venn Diagram. Set A is the synsets of hotel and set B is the synonyms of building. Set C is the result of the similarity distance between hotel and building. The yellow color shows the results of the intersection and the union between sets A, B, and C.

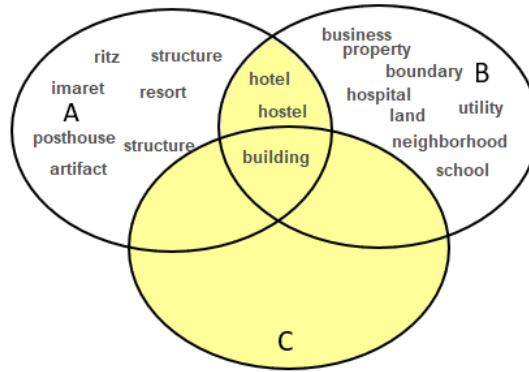


Figure 3.8: Results of "hotel" and "building" keywords in Venn Diagram

3.7. Scenarios algorithms

In this section, we describe how the scenarios explained in the previous section are implemented in the Python codes. We explain the steps of the multilingual WordNet scenario that has encompassed three sub-algorithm that correspond with the baseline, multilingual, and WordNet scenarios. The first sub-algorithm is the multilingual WordNet scenario. The first step is to lemmatize each keyword (both query keyword and metadata attributes). Lemmatize keywords function gets keywords and returns the root form of the words and removes specific characters (e.g., plural suffixes) from the keyword(s), and makes data compatible with the WordNet data type. Then, the synonyms function gets keywords in metadata keyword sets and a query keyword to compute and return the synonyms lists. The first four synonyms are considered relevant and the rest are filtered out to avoid noise in the result. Hypernyms and hyponyms are respectively got a query keyword and return super name and sub-name lists. Next, synonyms, hypernyms, and hyponyms resulted from the query keyword are combined to create set A. Set B is the synonyms list of metadata. To compute similarity, function C gets a query keyword, and the metadata keyword set as inputs and returns a keyword list above 0.69 similarity score. Lastly, the semantic overlay function is used. This function gets A, B, and C sets as inputs and computes the intersection between set A and set B, and the results are combined with set C and, finally, it returns semantic keyword list in English as the output of WordNet.

The second sub-algorithm is the translation system that uses Google translate API to translate the semantic keywords and query keywords in English and returns semantic keywords in Dutch. Please note that WordNet generates semantic keywords in English and the semantic overlay provides context for the translation system. Finally, Google translate API uses the context to produce compatible semantic keywords with metadata in Dutch and handle ambiguity and polysemous.

The last sub-algorithm is the baseline (query over metadata). These codes allow connection to metadata and search for datasets using SPARQL. The first step is to set a query over RDF metadata using the RDF library. The configuration of the query covers the name, keywords, about, and description of metadata. The next step is comparing strings using the partial ratio to measure the distance similarity ratio between query keyword and metadata strings. If the similarity score is equal to 100, the corresponding URIs are extracted from the metadata. The partial ratio in the fuzzy match string reduces the number of mismatch keywords for substrings and irrelevant datasets compare to other fuzzy string matching functions, such as standard Levenshtein distance similarity ratio or fuzz token functions. The output of this step is a list of URLs and the number of links. The output function provides a

CSV file for each keyword that consists of links, number of links, query keywords in English and Dutch. Lastly, the Python timer function is used to monitor the performance of retrieval results in seconds. Algorithm 1 is a pseudocode description that represents the algorithm steps as follow:

Algorithm 1 Multilingual WordNet Scenario

```

Set Inputs: user keyword, keywords extracted from metadata
Set Outputs: List of relevant and irrelevant dataset, file name, number of links, and links
Begin # WordNet Sub-algorithm
Function lemmatize keywords
    pass: keywords
    Lem keyword = lemmatizater (keywords)
    pass out: lemmatize keywords
End function
Function synonyms
    pass: lemmatize keywords
    compute wordnet synsets
    pass out: synonyms
End function
Function hypernyms
    pass: keywords
    compute keywords hypernyms
    pass out hypernyms
End function
Function hyponyms
    pass: keywords
    compute keywords hyponyms
    pass out hyponyms
End function
Set A = synonyms U hypernyms U hyponyms #query keywords
Set B = synonyms #metadata keywords
Function Set C
    pass: lemmatize query keyword, lemmatize metadata keywords
    compute wup path similarity
    pass out: keyword list above 0.69 similarity
End function
Function semantic overlay
    pass: Set A, B, C
    compute semantic overlay between  $A \cap B \cup C$ 
    pass out: Semantic keywords
End function
End# WordNet Sub-algorithm

Begin # translate system
Function translate
    pass: Semantic keywords
    language detection and translation
    pass out: translate Semantic keywords
End function
End# google translate

Begin #Baselin/ Query over metadata
Function query
    pass: translate Semantic keywords, RDF metadata
    set SPARQL statement
    compare and score strings
    if score == 100:
    pass out: URLs, number of links
End function
Function output
    pass: links, number of links, query keyword in English and Dutch, List of relevant and irrelevant dataset
    write inputs
    pass out: SCV file
End function
Function time
    set timer
    pass out: second (start-end)
End function
End# query over metadata

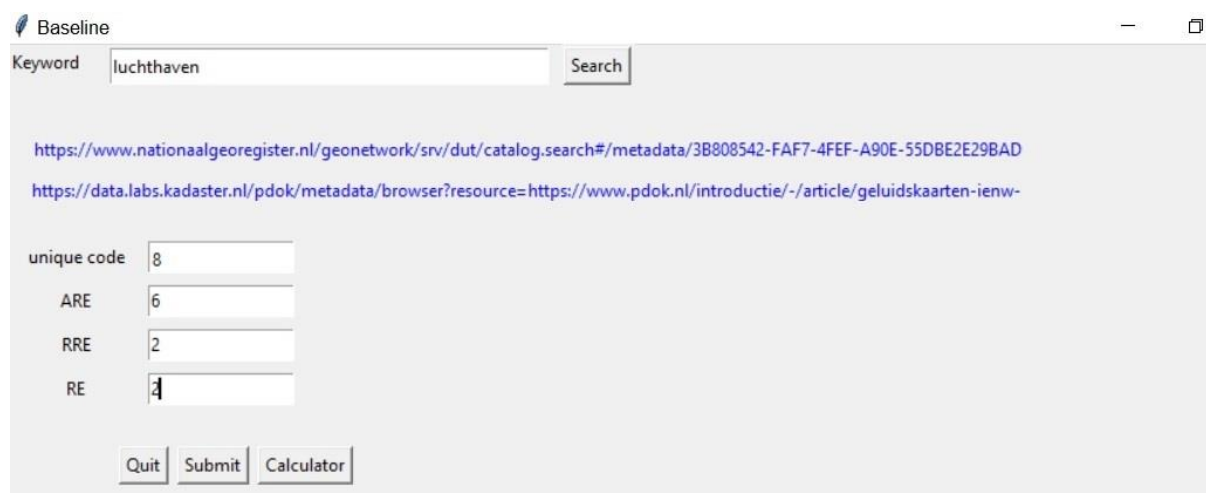
```

3.8. The fifth phase: IR evaluation metrics

The evaluation task is to extract a list of answers from each query keyword and record them in the corresponding files. And, finally, compare the results of scenarios (cf. Phases 4 and 6). The evaluation takes place with respect to the same SPARQL setting in the baseline, to ensure comparable results. Figure 3.9 represents the overview of the prototype search platform for query keywords in the baseline scenario. Each query keyword has a unique code, the total number of relevant answers for keyword 8 in the gold standard (ARE), the total retrieved links (RE) for query 8, and the total number of relevant answers for keyword 8 (RRE). The dataset for each keyword is submitted in the baseline retrieval set document. Then, three common evaluation metrics in IR for unranked documents are computed to evaluate the retrieval performance. These indices are the standard recall, precision, and F-measure (cf. Section 2.5). Using these indices, we aim to answer these questions:

Recall: "What ratio of relevant metadata is retrieved for each keyword? "

Precision: "What ratio of the retrieved metadata by the system is relevant to the query keywords? "



unique code	8
ARE	6
RRE	2
RE	4

Figure 3.9: Overview of a prototype platform

3.9. The sixth phase: Online tools scenarios

In addition to the mentioned scenarios in section 3.6, RESTful API on PDOK, NGR search engine, and linked data search engine are used to investigate the effectiveness and the efficiency of the proposed approach on online services. Therefore, three scenarios are defined for the RESTful API service including online retrieval baseline, online retrieval multilingual, and online retrieval WordNet. In addition to the mentioned platforms, NGR, and keywords search engines are used to examine the query keyword datasets (both Dutch and English). In this experiment, the search engine results are recorded with the respective source of metadata in the gold standard. For example, if the metadata source is RDF, the query keyword results are compared with the keyword search engine on PDOK. If the source is XML metadata, it is compared with the NGR search engine. And if the retrieval datasets have resulted from both sources, the combinations of search engines are considered (cf. Section 3.4). This discrimination between two sources is carried out manually and by comparing the results with RDF triple on the local machine.

Figures 3.10, 3.11, and 3.12 demonstrate the main components of the online baseline, multilingual, and multilingual WordNet scenarios. For the online baseline scenario, Dutch query keywords in the gold standard are used to query over metadata on PDOK using RESTful API and Python code. The retrieval results of this step are recorded on the online baseline retrieval set file. For the online multilingual and WordNet, English query keywords in the gold standard are used to fire query over metadata. And the retrieval results are documented in the online multilingual retrieval set file, and the online multilingual WordNet retrieval set file, respectively. Please note that the same algorithms for the translation and query expansion in the local scenarios, query keyword datasets, SPARQL setting, and gold standard are reused for the online scenarios. The only difference between the online tools scenarios with the local scenarios is in the baseline scenario (cf. Figure 3.10). In this scenario, RESTful API uses the SPARQL endpoint to return the datasets stored on the cloud. SPARQL setting covers the name, about, keywords, and description of metadata.

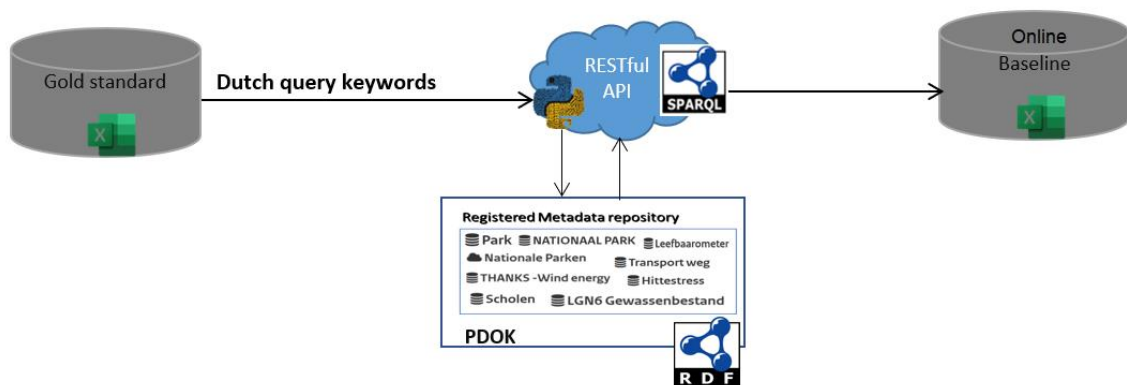


Figure 3.10: Components of the online baseline scenario

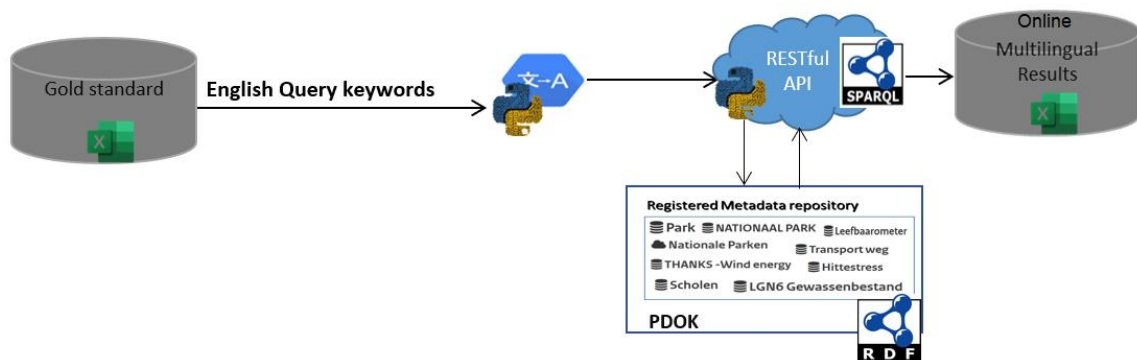


Figure 3.11: Components of the online multilingual scenario

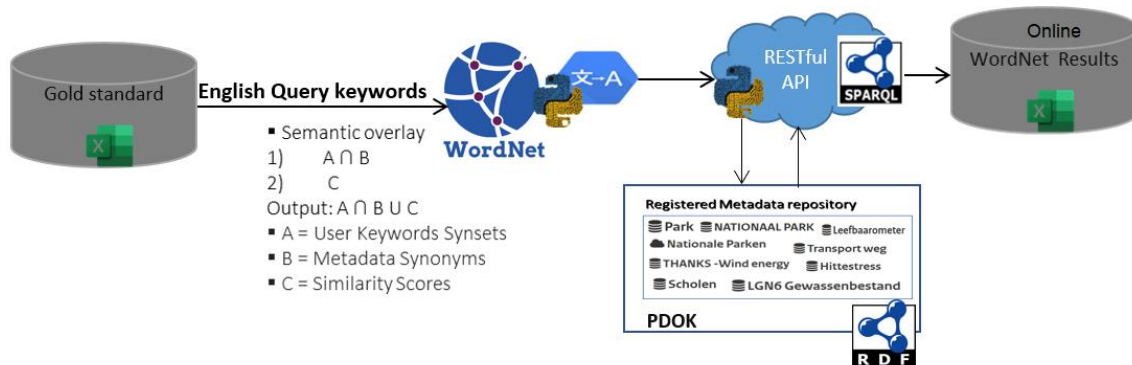


Figure 3.12: Components of the online multilingual WordNet scenario

Figures 3.13 and 3.14 represent the components of the online Dutch search engines and the English search engines scenarios, respectively. As can be seen in Figure 3.13, Dutch query keywords in the gold standard are used in both keyword search engine and NGR search engine, and results are recorded in the dataset (Dutch search engines retrieval set file). Figure 3.14 shows the components of the online English search engines scenario. English query keywords, in the gold standard, query over both keyword search engine and NGR search engine, and results are recorded in the English search engines retrieval set file. Finally, IR metrics explained in section 3.8 are used to calculate precision, recall, and F-measure for each query keyword in the online scenarios.

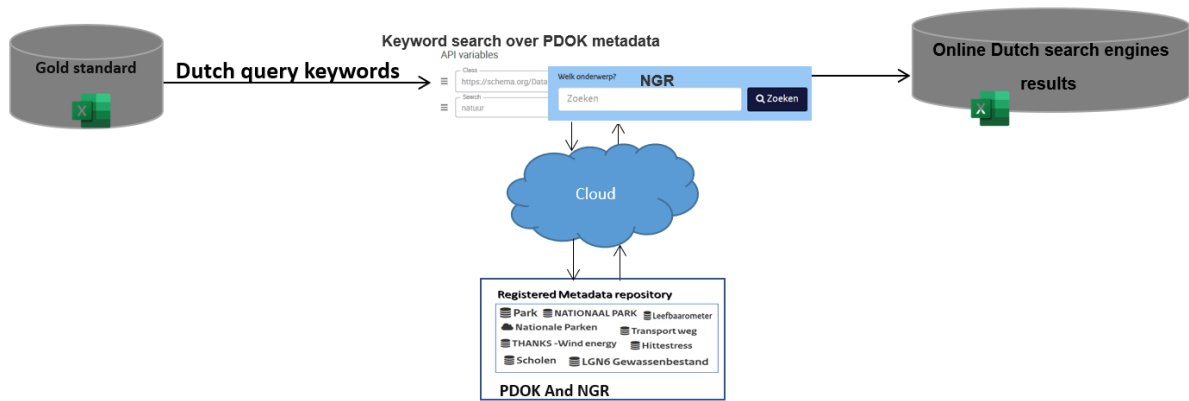


Figure 3.13: Components of the online Dutch search engines scenario

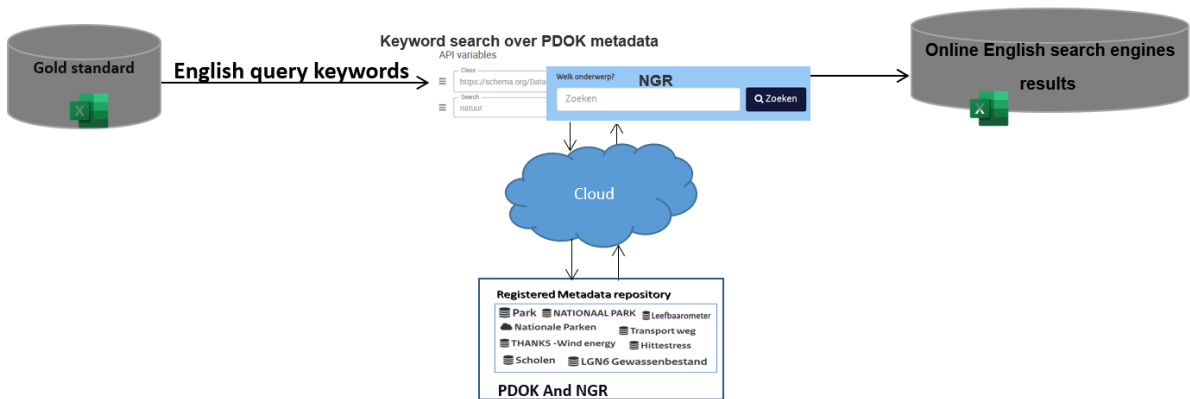


Figure 3.14: Components of the online English search engines scenario

4. Results and discussion

This chapter provides detailed information about the results of the steps in the methodology. The results of each section are described together with the discussions on the results.

4.1. Results of the local scenarios

This section represents the results of phase 4 in the methodology. First, we report on the results of each scenario separately to provide more detailed information. Then, we discuss the results.

4.1.1. Results of the local baseline scenario

In this section, we, first, describe the result of the baseline scenario compared to the gold standard. Then, we report on the results of IR metrics in this scenario.

Overall, the baseline scenario generated slightly fewer retrieval results, which are the outcome of different search methods (manual vs. automatic), than the gold standard (cf. Sections 3.5 and 3.6.1). More precisely, the gold standard embraces 167 query keywords in natural language, and 54% of the keywords (90 keywords) matched with metadata, whereas 86 out of 167 (51%) query keywords have resulted in the local baseline scenario.

Table 4.1 shows the results of precision, recall, and F-measure on 167 search tasks. The results of IR indices have been categorized into three classes. *Class 0* is query keywords with empty results, *class 1* represents the complete or ideal retrieval results, and the third class (*class 0 - 1*) indicates the retrieval results between zero and one. *The number* column indicates the total number of queries in each class. *Percent* column shows the percentage of the total number in each class. As a general trend, IR indices show large numbers of query keywords with ideal retrieval results and empty results. 48% (81 keywords) of queries have no answers. 30% of query keywords (50 keywords) follow the ideal trade-off i.e. high precision and recall. 22% of query keywords (36 keywords) show the inverse trade-off between recall and precision, 12 (7%) out of 36 query keywords have higher recall than precision, and 24 (15%) out of 36 keywords represent higher precision than recall.

Category	Recall		Precision		F-measure	
	Number	Precent	Number	Precent	Number	Precent
0	81	48%	81	48%	81	48%
0 - 1	25	15%	18	11%	36	22%
1	61	37%	68	41%	50	30%

Table 4.1: Recall, precision, and F-measure of baseline scenario on 167 search tasks

Please note that if all relevant answers are retrieved, recall is recorded 1, and F-measure represents the trade-off between precision and recall (cf. Section 3.8). Therefore, when F-measure, in class 1, is 30%, it indicates that 30% of query keywords have equal precision and recall. The inverse trade-off seen in the class of 0 - 1 represents high recall and low precision or vice versa. Moreover, in class 1, higher precision than recall demonstrates that the baseline scenario returns few relevant documents from the metadata repository. While the lower precision than recall shown in the class of 0 - 1 indicates that the total number of irrelevant answers has increased.

4.1.1.1. Discussion on the results of the local baseline scenario

In conclusion, the results of the local baseline scenario are encouraging. The results provide insight into the number of query keywords covered by the local baseline. As expected, the baseline scenario shows slightly fewer retrieval results than the gold standard (3%). Although this scenario has experienced a 3% reduction in retrieving relevant answers, this result is acceptable to make this scenario a benchmark for the evaluation and a query platform for other scenarios. So, 44%, 48%, and 46% are the base of recall, precision, and f-measure, respectively, for the other scenarios (cf. Table 4.6). Moreover, several limitations were recognized in the baseline scenario that harms the retrieval results. These limitations are metadata content, string matching algorithm, and metadata keywords.

The first problem is the metadata contents. More precisely, as mentioned before, the SPARQL setting covers keywords, about, name, and description. Using the description is a double-edged sword. On the one hand, not all keywords used in the descriptions are part of desired keywords. As a result, it increases the number of irrelevant datasets. On the other hand, removing the description from the SPARQL setting reduces the number of relevant datasets. The next problem is when the length of semantic keywords is shorter than the length of strings in the metadata. Therefore, keywords are matched with the part of strings and increase the number of irrelevant datasets (e.g., sport and transport). The last issue is related to the completeness and classification aspect of metadata. For example, the "scholen" (schools) metadata consists of "schools", "transport", and "utility" keywords. If a user searches for "transport" or "utility", a retrieved dataset is schools. Therefore, the incompleteness and classification problems have hurt the results of some queries.

4.1.2. Results of the local multilingual scenario

In this section, we, first, report on the precision of the translation system. Then, the results of the IR indices are explained. As mentioned in sections 4.1.1, 3.5, and 3.6.1, the gold standard and the baseline scenario are two benchmarks used for the evaluation of this scenario. So, the results of the translation system are compared to the gold standard and the baseline. By this comparison, we aim to understand whether the translation results are reliable and have an adverse influence on the number of retrieval results in the multilingual scenario.

Overall, the total number of translated keywords matched with the gold standard is 150 (90%) out of 167. This scenario has experienced less number of retrieval compared to the baseline. The retrieval results are recorded for 81 queries (49% of query keywords) that show a 2% reduction compared to the baseline. The main reason for this reduction is in different translation methods. Table 4.2 represents the sample of the query keywords that have not matched with the results of the baseline scenario and gold standard. The *multilingual scenario* and *baseline scenario* columns show the results of automatic translation and manual translations, respectively. The results column indicates the retrieval results categorized into two classes: *few answers*, and *no answer*. *Few answers* specify less number of retrieval compared to the benchmarks. For example, both "vliegveld" and "luchthaven" are keywords used in metadata for "airport", but the "vliegveld" retrieves more datasets than "luchthaven" since more metadata have been labeled "vliegveld". *No answer* shows the empty link for the query results.

The main reason for the difference between the benchmarks and this scenario is the polysemous (having multiple meanings). Google translate API often uses the frequency of translation and only returns one synonym for each term. For example, the keyword "plant" is

expected to be translated into "plant" as "a living organism" and "factory"; yet the Google translate API returns only "fabriek" (factory).

Keyword	Baseline scenario	Multilingual scenario	Results
airport	vliegveld	Luchthaven	Few answer
accident	ongeval	Ongeluk	No answer
conservation	beschermd	Behoud	No answer
fire station	Brandweerpost	Brandweerkazerne	No answer
houses	woningen	Huizen	Few answer
noise	geluid	Lawaai	Few answer
neighbor	buur	buurman	No answer
plant	plant	fabriek	No answer
water wells	waterput	waterbronnen	No answer
forest	bos	woud	No answer

Table 4.2: Google translate interface vs. the Google translate API

Table 4.3 illustrates the results of IR metrics in this scenario. As can be seen, a significant number of query keywords have no answer that covers 86 queries (51%). 27% of query keywords have equal high precision and recall in F-measure. That indicates a complete trade-off between precision and recall. 22% of query keywords (36 keywords) show the inverse trade-off between recall and precision; 11 out of 36 (7%) query keywords have higher recall than precision, and 15% of query keywords have higher precision than recall. This scenario, like the baseline, has experienced higher precision than recall. This indicates that fewer relevant documents are retrieved compared to the gold standard.

Category	Recall		Precision		F-measure	
	Number	Precent	Number	Precent	Number	Precent
0	86	51%	86	51%	86	51%
0 - 1	26	16%	17	10%	36	22%
1	55	33%	64	39%	45	27%

Table 4.3: Recall, precision, and F-measure of multilingual scenario on 167 search tasks

4.1.2.1. Discussion on results of the local multilingual scenario

The results gained in this scenario prove the high precision (86%) of near languages (Dutch, English) argued by Sequeira et al., 2020. Our approach shows 90% precision for the translation compared to the gold standard. Furthermore, 45 (27% of query keywords) out of 81 query keywords have an ideal trade-off between precision and recall (i.e., a 3% reduction compared to the baseline). The numbers of query keywords with no answer have increased. The results of recall and precision have declined in this scenario compared to the local baseline, since polysemous and ambiguity cause difficulty for the translation system. However, in many cases (90%), the frequency of translation handles the lexical ambiguities and polysemous. In conclusion, the automatic translation system can help the retrieval systems to overcome the language barrier with acceptable precision.

4.1.3. Results of the local WordNet scenario

In this section, first, the results of multilingual WordNet are compared to the gold standard and the aforementioned scenarios. Then, the results of evaluation metrics are described. In the next section, the results of this scenario are discussed.

Table 4.4 represents the overview of query expansion in the WordNet scenario compared to the gold standard and other scenarios. *No. queries* and *Amount of answer* columns indicate the total number of queries with the retrieval results and the retrieval percentage, respectively. In general, the retrieval results are 69% of query keywords which show a 15%, 18%, and 20% increase compared to the gold standard, the baseline, and multilingual scenarios, respectively. 31% of query keywords (51 keywords) are not retrieved any answer. 18% of keywords are not available in the WordNet. More precisely, WordNet suffers from word limitation and cannot support all geographic keywords, therefore, 30 keywords (18%) are not available in WordNet. 11 out of 30 keywords are matched with metadata based on the gold standard and the baseline scenario. And 19 out of 30 keywords are required to be expanded. Besides, we recovered 8 out of 11 keywords without any query expansion and only using the translation system over metadata (cf. Section 4.1.2). Lastly, 18% of keywords did not retrieve any answer. 2% of queries did not retrieve any answer due to wrong translation.

Scenarios	No. queries	Amount of answer
Gold standard	90	54%
Baseline scenario	86	51%
Multilingual scenario	81	49%
WordNet scenario	116	69%

Table 4.4: The results of WordNet

Table 4.5 provides information about the precision, recall, F-measure in the multilingual WordNet scenario. Generally, there has been a noticeable improvement in the number of recalls. 62% of queries (103 keywords) have produced complete recall. However, the high recall has not led to high precision. Only 23% of query keywords have experienced high precision, whereas a significant amount of retrieval answers (47%) has witnessed the precision between 0 and 1. Moreover, 49% of query results show an inverse trade-off between precision and recall, in particular, high recall (45%) and low precision (4%), only 21% of queries represent the ideal trade-off. 58 out of 81 query keywords have a high F-measure above 0.5 in the class 0-1.

Category	Recall		Precision		F-measure	
	Number	Precent	Number	Precent	Number	Precent
0	51	30%	51	30%	51	30%
0 - 1	13	8%	78	47%	81	49%
1	103	62%	38	23%	35	21%

Table 4.5: Recall, precision, and F-measure of WordNet scenario on 167 search tasks

4.1.3.1. Discussion on results of the local WordNet scenario

The results of the local WordNet are promising in terms of the reliability and effectiveness of the algorithm. The results of the WordNet scenario have increased by 15% and 18% compared to the gold standard and the local baseline scenario. This indicates that this scenario covers a wide range of query keywords with retrieval results. Besides, an 18% reduction has occurred in class 0 compared to the baseline. The inverse trade-off has increased between 0 - 1 in this scenario, in particular high recall and low precision.

Moreover, in this study, we used hierarchical relationships (hyponyms, hypernyms) and similarity scores to deal with word sense ambiguity problems by creating the context for the semantic keywords. This approach was helpful to improve the results. A similar method was proposed by Aouicha et al., 2018. They presented an approach for word sense ambiguity problems using a relatedness score. They considered the synsets surrounding the ambiguous word considering the hierarchical relationships (hyponyms, hypernyms) in wordnet. Moreover, additional nouns were taken from the glosses of synsets in WordNet and Wiktionary. The results showed improvement by incorporating Wiktionary. This indicates that more semantic keywords can be helpful to increase retrieval results and decline the lexical ambiguities.

Furthermore, several reasons contribute to this high recall and low precision including, 1) string matching algorithm, 2) metadata keywords, 3) metadata completeness and classification, 4) results of the translation system and 5) the relatedness of semantic keywords 6) word limitation. String matching algorithm, metadata keywords, and metadata completeness and classification were addressed in the local baseline scenario (cf. Section 4.1.1.1). This section explores the last three problems.

The most important issue is word limitations in WordNet. More precisely, WordNet does not support all geographic terms, and 18% of query keywords are affected by the word limitation. For example, some query keywords (such as "land use") are not available in the WordNet. Another issue is related to the similarity function. Although the outputs of each function are limited to return the relevant keywords, in some cases, the less relevant semantic keywords have a higher score than the relevant ones. For example, the similarity score between "school" and "building" is 0.13, whereas this score for "building" and "road" is 0.7.

Another difficulty that often keeps the retrieval results from achieving 100% precision is the polysemous and ambiguity issues in Google translate API. These problems are minor for the WordNet scenario compared to the multilingual scenario. As mentioned in the multilingual scenario, 11 keywords are translated differently from the original keywords recorded in the gold standard. And they had less amount of retrieval or no answers results(cf. Table 4.2). In this scenario, Google translate API deals with the ambiguity and the polysemous problems, while WordNet produces more semantic keywords to provide context around keywords. As a result, only 4 keywords trigger no result in WordNet (e.g., "accident" and "conservation"). And 3 out of 11 keywords are not available in the WordNet. Although these issues are minor for the WordNet scenario compared to the multilingual scenario, this problem keeps the results from achieving high precision in some retrieval results. Lastly, although we used the semantic overlay to reduce the number of irrelevant datasets, the WordNet algorithm sometimes returns the less relevant dataset. For example, the results of the semantic keywords for "tornado" consist of "wind" and "flood" datasets. Although flood datasets are the irrelevant dataset, it is related to "tornado".

4.2. Final results of the local scenarios

This section provides insight by comparing all results to understand which NLP techniques are effective and efficient, and more suitable for the proposed corpus. Table 4.6 represents the IR indices results for each scenario, listing the average precision, recall, and F-measure values for query keywords with retrieval results. The *Total column* indicates the total number of query keywords with retrieval results. The *Correct answers* column specifies the total number of relevant links, and the *False answers* column states the total number of irrelevant links. The *avg.R. Time* column shows the average response time for queries.

As can be seen, in the local multilingual scenario, the recall and precision, and F-measure have decreased 4%, 3%, and 4% compared to the local baseline. However, the local WordNet scenario shows opposite results, and the geo-data recall has enhanced 22% compared to the local baseline, and the precision represents a 1% improvement. Furthermore, in the local WordNet scenario, the total number of relevant answers, compared to the local baseline scenario, has improved about 3 times. On the other hand, the expected implication is the total number of irrelevant links that increased about 10 times. In addition, the local multilingual scenario shows fewer retrieval results in both correct answers and false answers (cf. Table 4.6) compared to the local baseline scenario.

Moreover, the average response times are 3 and 4 seconds in the local baseline and the local multilingual scenarios, respectively, whereas the average response time is 11 seconds in the local multilingual WordNet. This indicates that the computation cost, compared to the local baseline, has increased about 4 times.

Scenarios	Total	Correct answers	False answers	Avg.Recall	Avg.Precision	Avg.F-measure	Avg.R. Time
Baseline	86	599	141	44%	48%	46%	3 second
Multilingual	81	514	137	40%	45%	42%	4 second
WordNet	116	1642	1363	66%	49%	56%	11 second

Table 4.6: Results of recall, precision, and F-measure

4.2.1. Discussion on final results of the local scenarios

This section discussed the results of the local scenarios. The local baseline and multilingual scenarios produce higher precision than recall. This indicates that the scenarios are more "careful" in retrieving datasets. Therefore, they retrieve answer sets that contain a higher proportion of relevant answers compared to total retrieval answers. Yet, they have missed more of the relevant items in the metadata repository (RDF). The WordNet scenario returns high recall than precision. It means although the local WordNet scenario is good at retrieving relevant metadata from the metadata repository; it is less "careful" in retrieving datasets.

Overall, the results of 167 queries directed at the local scenario indicate that the local WordNet scenario is the most effective approach and presented the best performance based on IR metrics. This scenario has enhanced the precision, recall, F-measure of geo-datasets by 1%, 22%, and 10%, respectively. Precision indicates that the local WordNet returned sets containing a higher proportion of relevant metadata compared to the baseline. Recall shows

that the WordNet scenario also was good at handling the relevant metadata in document collection (RDF) at large compared to the baseline. F-measure shows an increase in the inverse trade-off between recall and precision compared to complete trade-off. Moreover, the local multilingual scenario results show a 4%, 3%, and 4% reduction in the recall, precision, and F-measure compared to the baseline scenario. The results indicate that the translation system can handle the language barrier. Furthermore, the integration of WordNet and Google translate can effectively deal with the ambiguity of query keywords in both languages.

4.3. Results of the online tools scenarios

In this section, we aim to study the effectiveness of the proposed approach using online services. As mentioned in section 3.9, RESTful API, on the PDOK, is selected to examine the algorithm over the online tools. Besides, the search engines are used to examine to what extend query keywords are supported by the search engines. In this section, we, first, compare the results of the online baseline and search engines compared to the gold standard. Then, the online multilingual and the online multilingual Wordnet scenarios are compared with the online baseline.

Generally, the category of the online scenarios has resulted in less amount of retrieval results compared to the gold standard. This reduction is 17% in the online baseline scenario. In this scenario, 62 out of 167 (37%) query keywords have returned answers using metadata on the cloud. 88 query keywords have retrieved the results using Dutch search engines and represent a 2% decline in retrieval results compared to the gold standard. This result is much lower in the English search engines scenario. The English search engines have witnessed a 41% reduction compared to the gold standard. The main reason for the fewer retrieval results in the online scenario compared to the gold standard is that they have fewer metadata search options. As explained in section 3.2, metadata for the local scenarios is enriched with extra metadata. As a result, metadata on the local machine covers more datasets and is more compatible with the query keywords.

Table 4.7 represents the results of query keywords datasets for each retrieval scenario, listing the average precision, recall, and F-measure values for query keywords in each scenario. The *Total* column indicates the total number of queries with retrieval results, the *Correct answers* column specifies the total number of relevant answers, and the *False answers* column states the total number of irrelevant answers. The *avg.R.Time* column shows the average response time for queries.

Online Scenarios	Total	Correct answers	False answers	Avg.Recall	Avg.Precision	Avg.F-measure	Avg.R. Time
Online Baseline	62	677	274	31%	32%	32%	00.4 second
Online Multilingual	55	565	271	27%	28%	27%	00.5 second
Online WordNet	87	1561	3475	48%	30%	37%	6 second
English Search engines	25	81	12	12%	14%	12%	-
Dutch Search engines	88	633	11	50%	52%	51%	-

Table 4.7: Results of recall, precision, and F-measure in non-participant platforms

Moreover, as shown in Table 4.7, the results of recall and precision, and F-measure have decreased 4%, 4%, and 5% in the online multilingual scenario compared to the online baseline. On the other hand, although, in the online WordNet scenario, the recall and the F-

measure have enhanced by 17% and 5%, respectively, and precision has experienced a 4% reduction.

Furthermore, the online multilingual scenario shows fewer retrieval results for total correct answers compared to the online baseline (cf. Table 4.7). However, the total number of irrelevant links has increased about 12 times. In the online WordNet scenario, the total number of relevant answers has enhanced about 2 times compared to the online baseline scenario.

4.3.1. Discussion on results of the online tools scenarios

In conclusion, the online baseline and online multilingual scenarios have witnessed higher precision than recall. It indicates that these scenarios are more "careful" in retrieving datasets from the cloud. Hence, they retrieve answer sets containing a higher proportion of relevant metadata, but they have missed more relevant datasets in RDF metadata. The online WordNet scenario has returned high recall than precision; it means although the online WordNet scenario is good at retrieving relevant metadata from the metadata repository, it is less "careful" in retrieving datasets.

Overall, the results of 167 query keywords show that the online WordNet scenario was slightly better than other online scenarios and represented good performance based on IR metrics. This scenario has improved the recall and F-measure by 17% and 5%, respectively. The precision shows a 2% reduction compared to the baseline. So, the online WordNet returned sets containing a lower proportion of relevant metadata compared to the online baseline. The result of recall shows that the online WordNet scenario was good at dealing with the relevant metadata from the metadata repository compared to the online baseline. F-measure represents that the trade-off between recall and precision has slightly increased. Moreover, the results of the online multilingual scenario have decreased by 4%, 4%, and 5% in the recall, precision, and F-measure, compared to the online baseline scenario. Furthermore, the Dutch search engines show slightly fewer retrieval results compared to the gold standard, whereas there is a significant difference between the results of English search engines and the gold standard.

4.4. The online tools scenarios vs. the local scenarios

This section compares the results of the online tools and the local tools. With this comparison, first, we aim to understand which of the online services and the local scenarios are effective and efficient among the NLP techniques. So, in this section, we compare online services, in particular online services that use RESTful API and the same SPARQL setting and the same algorithm, with the local scenarios. Then, by comparing the online search engines with the corresponding local scenarios, we aim to understand to what extent the search engines are effective to support the query keywords.

As can be seen, in tables 4.6 and 4.7, the proposed approach, in the local scenarios, shows better results compared to the online scenarios. The results of the online baseline show a 13%, 16% decline in recall and precision compared to the local baseline. Moreover, recall and precision have decreased by 13% and 17%, in the online multilingual scenario, compared to the local scenario. Besides, the online WordNet has witnessed an 18% and 19% decline, in recall and precision, compared to the local WordNet.

Moreover, the total number of irrelevant answers, in the online scenarios, shows a significant growth (2 times) compared to the local scenarios. The main reasons for this

difference are the difference in metadata (cf. Section 3.4) and the string matching algorithm. Also, the mentioned online scenarios are much better than the local scenarios in terms of response time, and response times have improved compared to the local scenarios.

Besides, the English search engines have witnessed a significant reduction compared to the local multilingual scenario. Recall and precision have declined by 13% and 17%, respectively. Finally, in the Dutch search engines, we used Dutch query keywords (cf. Section 3.9) to help the search engines and retrieve the dataset. The retrieval results enhanced 6% and 8% in precision and recall compared to the local baseline.

4.4.1. Discussion on online tools scenario vs. local scenario

In this section, we discuss which NLP techniques are more effective and efficient. As mentioned before, the local scenarios show better results compared to the online scenarios. This indicates that the local scenarios are more effective in terms of IR indices. However, the results of response time demonstrate that the local scenarios are not efficient in terms of the performance of queries. As a result, the response time significantly increased.

Moreover, although the results of the local scenarios indicate that the local WordNet is much better than that of other local scenarios in terms of IR indices, the local WordNet is not efficient in terms of response time and represents lower performance. Moreover, in the local WordNet, the response time has increased 2 times compared to the online WordNet scenario. Additionally, although the English search engine, unlike the local multilingual scenario, cannot support the wide range of query keywords, the Dutch search engine was much better than that of the baseline scenario in terms of IR indices.

5. Conclusion

In this thesis, we proposed a hybrid approach to enhance the findability of geodata sources using NLP techniques. For this purpose, query keywords were extracted from GeoAnQu corpus, and metadata was gathered from the PDOK and NGR. A gold standard was published to be a benchmark for the evaluation. Three scenarios were defined to investigate NLP techniques on the local machine. The local baseline scenario was devised to examine the naive query over metadata using SPARQL language and fuzzy string matching. Google translate API was employed to inspect the cross-lingual issue by establishing a translation system. Finally, the multilingual WordNet platform was built on top of other platforms to handle the mismatch query. The WordNet platform is an example of a mature search platform in which terms were mapped in three dimensions to generate semantic keywords, and a heuristic algorithm was customized to capture the synsets of query keywords and the similarity between query keywords with metadata. The semantic overlay was used to have better control over the ambiguity and the reduction of retrieval results. Finally, the retrieval results of the scenarios were compared to the gold standard and the baseline. We also defined five online scenarios to compare the effectiveness and efficiency of the proposed corpus and NLP techniques applied in the local scenarios. Three scenarios were defined for the RESTful API with the same setting and algorithm used for the local scenario. In addition, two scenarios were defined for the search engines. Based on the quantitative analysis of 167 query results directed at different contexts, it can be concluded that the results of IR metrics in the local scenarios are encouraging. The best performance IR metrics were obtained through the local WordNet scenario that can retrieve 69% of query keywords expressed in natural language. This approach also promoted the precision and recall of geo-datasets by 1% and 22%, respectively, compared to the baseline scenario. The results indicate that the alternative queries recommended by the combination of WordNet and Google translate API could reflect the true intention of users. Although the results of the multilingual WordNet scenario were effective in terms of retrieval quality and IR metrics, the response time increased compared to the local baseline and the online WordNet scenario. This implies that the proposed multilingual WordNet is not efficient in terms of query performance.

These steps were taken to find an answer to the main research question, as presented in the introduction. The question is: *What types of NLP techniques can be used to improve the findability of geodata sources, and to what extent are NLP techniques efficient?*

The main research question has been split up into six sub-questions. Answers to these questions are formulated in the following section.

5.1. Research overview

In this section, first, the sub-questions are reviewed and answered. Then, limitation, future work, and contributions are explained.

- *How can NLP techniques be used to retrieve and search over metadata?* To answer this research question, query keywords were selected from a large corpus (GeoAnQu) that denote different geographic phenomena. Two national brokers (PDOK and NGR) were selected to harvest metadata using the search engines. RDF metadata was enriched by XML metadata and stored on the local machine. Furthermore, Python code was developed to execute SPARQL query and string matching against metadata for three scenarios. These scenarios were the baseline, multilingual, and WordNet. The baseline scenario was devised to examine the naive query over metadata. Google translate API was employed to investigate the cross-lingual issue by establishing a translation system.

Lastly, the multilingual WordNet scenario was built on top of the other platforms to be an example of a mature search platform in which semantic keywords were computed and translated. Moreover, 5 scenarios were defined for the online services on the brokers to retrieve datasets. The same algorithms, query keywords, and SPARQL settings used for the local scenarios were examined over the RESTful API using the SPARQL endpoint to search over metadata. Moreover, search engines were employed to search over metadata. Both English and Dutch keywords were investigated over the search engines and retrieval results were studied.

- *To what extent can multi-linguistics problems be handled using Google API?* In this research, the gold standard was used as the benchmark for the precision of the translation system and calculating recall. Moreover, the baseline was used as a base to compare IR indices. The total number of translated keywords matched with the gold standard was 90%. Furthermore, the results of recall, precision, and F-measure decreased by 4%, 3%, and 4%, respectively, compared to the baseline. The main problems are ambiguity and polysemous. As demonstrated in sections 3.6.2, Google translate API unlike the Google translate interface does not return different synonyms for each keyword. Indeed, Google translate API uses the frequency of translation and only returns one synonym for each term. Therefore, it cannot completely handle ambiguity and polysemous of keywords. Although the multilingual scenario produced lower results compared to the baseline, it provided 90% precision compared to the gold standard. It indicates that the automatic language translator can help the system and users to overcome the language barrier.
- *To what extent WordNet is effective for the query expansion method?* The results of sections 4.1.3 and 4.1.3.1 showed the proposed query expansion effectively enhanced precision and recall by 1% and 22% compared to the baseline. The high recall indicates that this scenario returns higher documents compared to the gold standard and the baseline scenario. On the other hand, low precision demonstrates that the number of irrelevant answers has increased. In this scenario, 69% of query keywords (116 keywords) had answers which increased by 15%, 18%, and 20% compared to the gold standard, the baseline, and multilingual scenarios, respectively. Moreover, in the multilingual WordNet scenario, the multi-linguistics problems were handled slightly well for the semantic keywords. The semantic keywords allowed the translation system to deal with ambiguity by providing a context around the query keyword. So, only 2% of semantic keywords did not lead to any answer. This indicates that our approach is effective. However, this scenario suffers from several limitations that keep the results from achieving 100% precision and recall. These limitations are unavailable geographic keywords in the WordNet, less relevant semantic keywords, and ambiguity and polysemous.
- *How much does the result of keyword expansion promote retrieval quality?* To answer this question, three IR indices were used. First, the results of the baseline were compared to the gold standard results. Then, the results of multilingual and WordNet scenarios were compared to the baseline using the IR indices. Overall, the baseline scenario generated 3% fewer retrieval results compared to the gold standard. The results of the baseline showed 4% higher precision than recall. In the multilingual scenario, IR metrics decreased by 4%, 3%, and 4% compared to the baseline scenario. This scenario experienced 5% higher precision than recall. The results of this scenario showed that this scenario could not enhance retrieval results. However, in the multilingual WordNet scenario, the results of recall, precision, and F-measure improved by 12%, 1%, and 10%,

respectively, compared to the baseline scenario. The WordNet scenario witnessed higher recall than precision. It shows that the multilingual WordNet scenario retrieves a high number of documents. The total number of relevant documents was 3 times higher than the baseline, and the total number of irrelevant documents increased 9 times compared to the baseline.

- *Which query expansion method is more suitable for the proposed corpus?* In chapter 4, the results of IR metrics showed us our approach in the multilingual WordNet is more appropriate for the GeoAnQu corpus. This scenario also covered a wide range of query keywords. The WordNet scenario could return answers for 69% of query keywords. This approach also promoted the precision and recall of geo-datasets by 1% and 22%, respectively, compared to the baseline scenario.
- *To what extent the proposed corpus and the NLP techniques are effective and efficient for the online services on the infrastructures?* In section 4.4, the five scenarios were evaluated by IR metrics. Then, they were compared to the local scenarios to investigate the effectiveness of the proposed corpus and the NLP techniques for the online infrastructure. The results of this section showed that all online scenarios have lower retrieval results compared to the corresponding local scenarios. The main reason for the fewer retrieval results in the online scenario compared to the local scenario is that they have fewer metadata search options. As explained in section 3.2, metadata, in the local scenarios, was enriched with extra metadata. Furthermore, although the English search engines experienced lower results, the Dutch search engines showed higher retrieval results compared to the local baseline. On the other hand, the results of response time demonstrated that the online scenarios are efficient in terms of the performance of queries.

5.2. Limitation

The proposed methodology is subject to several limitations and the results suggest that there is room for improvement:

- In the local WordNet scenario, the proposed synsets and similarity approaches could not entirely show the relation between two keywords with high similarity and relatedness. Back to the ambiguity in "green areas" explained in the introduction, in this work, our proposed approach could show only the relatedness between "green" and "park" using synonyms in hierarchical relationships. However, it could not entirely show the relatedness between "green" and other datasets (e.g., trees and forests). This is also true for "animal" and "fauna" or "crape myrtle" and "flora". More precisely, although "animal" and "crape myrtle" are related to "fauna" and "flora", "fauna" and "flora" are not defined in the hierarchal relationship of "animal" and "crape myrtle". The same problem can be seen where two similar and related keywords are expected to possess a higher score. For instance, the result of the semantic keyword for "library" is "building" and the similarity score between the two keywords is 0.75. However, the results of the similarity score between "primary school" and "building" are 0.13. As a result, this approach may not completely address the relationship between keywords.
- As mentioned in section 2.4.1.2, Ezzikouri et al., 2019 introduced a similarity score based on the set theory between synonyms and keywords used in the gloss of two words. In this research, we only used synonyms and similarity scores of keywords because using gloss keywords and the relatedness score can be problematic and increase the number of

irrelevant geo-data retrieval. More precisely, in many cases, keywords used in gloss may increase relatedness between two less relevant keywords. This approach may be helpful to separate man-made from natural features. Table 5.1 gives information about the short definition of sample keywords and the results of tokenization and POS tags in WordNet. Although the "tree" is related to "park" and "national park", it shows better relatedness to "forest".

Keywords	Definitions(gloss)	Keyword extraction results
Forest	The trees and other plants in a large densely wooded area.	['trees', 'plants', 'area']
National park	A tract of land declared by the national government to be public property.	['tract', 'land', 'government', 'property']
Park	A large area of land preserved in its natural state as public property.	['area', 'land', 'state', 'property']
Plant	Buildings for carrying on industrial labor.	['Buildings', 'labor']
Tree	A tall perennial woody plant having a main trunk and branches forming a distinct elevated crown.	['woody', 'plant', 'trunk', 'branches', 'crown']
Bank	Sloping land (especially the slope beside a body of water).	['land', 'slope', 'body', 'water']

Table 5.1: Result of keyword extraction from gloss

- The ambiguity and the polysemous of keywords are problems that cannot completely handle using only WordNet. This is also true for Google translate API; since Google translate API uses the frequency of translation (cf. Sections 4.1.2.1 and 4.1.3.1).
- To address the above-mentioned issues, we tried to incorporate FrameNet and Dutch WordNet and compute semantic keywords in FrameNet, yet most geographic terms were not available in FrameNet. Dutch WordNet (Postma et al, 2016) also covers a limited number of words, since it has been built on top of WordNet V1. Besides, one file (OpenDutchWordnet.py) was not available to work with this data source. As a result, these tools were excluded from this research.
- Another problem that reduced precision is the completeness and classification of metadata. For example, schools metadata contains "schools", "transport", and "utility" keywords. So, when a user searches for transport, one of the search results is school.

5.3. Future work and recommendations:

Concerning the limitations of this research the following future works are proposed:

- Using other online data sources and NLP tools (e.g., ConceptNet, Wiktionary, or Dutch spacy) could be a logical and compelling step in future research to address the above-mentioned limitations. Google translate API may deal with the ambiguity and the polysemous of keywords, while online data sources offer more relevant semantic keywords. Besides, these data sources may be useful to find semantic keywords for unavailable keywords in WordNet and show a better relationship between two related words.
- Furthermore, in this work, we only used partial ratio fuzzy matching to match the strings; however, other frameworks and functions can be beneficial. For example, PolyFuzz uses different fuzzy string matching techniques as a framework, such as Levenshtein distance, TF-IDF character-based, and n-gram methods together. This framework can be customized to model fuzzy string matching. These features make it valuable to be examined in the future.

- Lastly, two more questions were raised during this work, which might be exciting avenues for future work. First, how can the retrieval results of RDF metadata be weighted and ranked to retrieve more relevant datasets in the first n results? How do the NLP techniques improve the findability of geodata sources using ontology?

5.4. Contributions

The main motivation for investigating NLP approaches was to improve the findability of geodata sources and evaluate the effectiveness of the proposed methodology. The following points are the main contributions of this research:

- ✓ No previous study has been carried out using a large corpus for geo-information retrieval on SDIs.
- ✓ No previous study was performed to show the effectiveness of retrieval results using IR metrics on the SDIs. In this study, we only focused on IR metrics for unranked documents.
- ✓ There are a few studies to address ambiguity and the polysemous using similarity relatedness score for geographic phenomena and concepts.

6. References

- Aditya, T., & Kraak, M. J. (2007). A search interface for an SDI: implementation and evaluation of metadata visualization strategies. *Transactions in GIS*, 11(3), 413-435.
- AlMousa, M., Benlamri, R., & Khoury, R. (2020). Exploiting Non-Taxonomic Relations for Measuring Semantic Similarity and Relatedness in WordNet. *arXiv preprint arXiv:2006.12106*.
- Aouicha, M. B., Taieb, M. A. H., & Marai, H. I. (2018). Wordnet and wiktionary-based approach for word sense disambiguation. In *Transactions on Computational Collective Intelligence XXIX* (pp. 123-143). Springer, Cham.
- Banerjee, S., & Pedersen, T. (2003, August). Extended gloss overlaps as a measure of semantic relatedness. In *Ijcai* (Vol. 3, pp. 805-810).
- Ballatore, A., Bertolotto, M., & Wilson, D. C. (2015). A structural-lexical measure of semantic similarity for geo-knowledge graphs. *ISPRS International Journal of Geo-Information*, 4(2), 471-492.
- Ballatore, A., Bertolotto, M., & Wilson, D. C. (2013, April). Grounding linked open data in WordNet: The case of the OSM semantic network. In *International Symposium on Web and Wireless Geographical Information Systems* (pp. 1-15). Springer, Berlin, Heidelberg.
- Ballatore, A. (2013). *Semantics and Similarity for Crowdsourced Geospatial Data* (Doctoral dissertation, University College Dublin).
- Bao, J., Duan, N., Zhou, M., & Zhao, T. (2014, June). Knowledge-based question answering as machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 967-976).
- Brewster, C., Alani, H., Dasmahapatra, S., & Wilks, Y. (2004). Data driven ontology evaluation.
- Billen, R., Noguera-Iso, J., López-Pellicer, F. J., & Vilches-Blázquez, L. M. (2011). Ontologies in the Geographic Information sector. In *Ontologies in Urban Development Projects* (pp. 83-103). Springer, London.
- Bouscarrat, L., Bonnefoy, A., Capponi, C., & Ramisch, C. (2020). Multilingual enrichment of disease biomedical ontologies. *arXiv preprint arXiv:2004.03181*.
- Bucher, B., Tiainen, E., Brasch, T. E. V., Janssen, P., Kotzinos, D., Čeh, M., ... & Zhral, M. (2020). Conciliating perspectives from mapping agencies and Web of data on successful European SDIs: Toward a European geographic knowledge graph. *ISPRS International Journal of Geo-Information*, 9(2), 62.
- Buscaldi, D., Rosso, P., & Sanchis, E. (2006, September). A wordnet-based indexing technique for geographical information retrieval. In *Workshop of the Cross-Language*
- Cai, Z., Kalamatianos, G., Fakas, G. J., Mamoulis, N., & Papadias, D. (2020). Diversified spatial keyword search on RDF data. *The VLDB Journal*, 1-19.
- Ceri, S., Bozzon, A., Brambilla, M., Della Valle, E., Fraternali, P., & Quarteroni, S. (2013). An introduction to information retrieval. In *Web information retrieval* (pp. 3-11). Springer, Berlin, Heidelberg.
- Chen, Z., & Yang, Y. (2020). Semantic relatedness algorithm for keyword sets of geographic metadata. *Cartography and Geographic Information Science*, 47(2), 125-140.
- Chen, Z., Song, J., & Yang, Y. (2018). Similarity measurement of metadata of geospatial data: An artificial neural network approach. *ISPRS International Journal of Geo-Information*, 7(3), 90.
- Costa, T. S., Gottschalk, S., & Demidova, E. (2020). Event-QA: A Dataset for Event-Centric Question Answering over Knowledge Graphs. *arXiv preprint arXiv:2004.11861*.
- David, W. A. (2010). *GIS Tutorial 2: Spatial Analysis Workbook*.
- Dalianis, H. (2018). *Clinical text mining: Secondary use of electronic patient records*. Springer Nature.
- Degbelo, A., & Teka, B. B. (2019, November). Spatial search strategies for open government data: A systematic comparison. In *Proceedings of the 13th Workshop on Geographic Information Retrieval* (pp. 1-10).
- Deleger, L., Lingren, T., Ni, Y., Kaiser, M., Stoutenborough, L., Marsolo, K., ... & Solti, I. (2014). Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research. *Journal of biomedical informatics*, 50, 173-183.
- Diefenbach, D., Both, A., Singh, K., & Maret, P. (2020). Towards a question answering system over the semantic web. *Semantic Web*, (Preprint), 1-19.
- Elbedweihy, K., Wrigley, S. N., Ciravegna, F., & Zhang, Z. (2013, October). Using BabelNet in Bridging the Gap Between Natural Language Queries and Linked Data Concepts. In *NLP-DBPEDIA@ ISWC*.
- Espinoza, M., & Mena, E. (2007, June). Discovering web services using semantic keywords. In *2007 5th IEEE International Conference on Industrial Informatics (Vol. 2, pp. 725-730)*. IEEE.
- Ezzikouri, H., Madani, Y., Erritali, M., & Oukessou, M. (2019). A New Approach for Calculating Semantic Similarity between Words Using WordNet and Set Theory. *Procedia Computer Science*, 151, 1261-1265.
- Fathalla, S., Vahdati, S., Lange, C., & Auer, S. (2019, October). SEO: A scientific events data model. In *International semantic web conference* (pp. 79-95). Springer, Cham.

- Flank, S. (1998, August). A layered approach to NLP-based information retrieval. In 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1 (pp. 397-403).
- Florence, M. (2020). MLGrafViz: multilingual ontology visualization plug-in for protégé. *Computer Science and Information Technologies*, 2(1), 43-48.
- Folmer, E., Ronzhin, S., Van Hillegersberg, J., Beek, W., & Lemmens, R. (2020). Business Rationale for Linked Data at Governments: A Case Study at the Netherlands' Kadaster Data Platform. *IEEE Access*, 8, 70822-70835.
- Gong, Z., Cheang, C. W., & Hou, U. L. (2005, August). Web query expansion by WordNet. In *International Conference on Database and Expert Systems Applications* (pp. 166-175). Springer, Berlin, Heidelberg.
- Gupta, P., & Gupta, V. (2012). A survey of text question answering techniques. *International Journal of Computer Applications*, 53(4).
- Heywood, I., Cornelius, S., Carver, S. (2011): *An Introduction to Geographical Information Systems*. Pearson Education Limited, Harlow
- Hirst, G., & St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database*, 305, 305-332.
- Jijkoun, V., Mur, J., & de Rijke, M. (2004). Information extraction for question answering: Improving recall through syntactic patterns. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics* (pp. 1284-1290).
- Jones, C. B., & Purves, R. S. (2008). Geographical information retrieval. *International Journal of Geographical Information Science*, 22(3), 219-228
- Kadaster(2020.a) Elastic search engine. Retrieved from: <https://data.labs.kadaster.nl/pdok/metadata/search>.
- Kadaster(2020.b) Search engine for browser. Retrieved from: <https://data.labs.kadaster.nl/pdok/metadata/browser>.
- Kadaster(2020.c) Keyword Search. Retrieved from: <https://data.labs.kadaster.nl/wouter/-/queries/key-word-search/4>.
- Kadaster(2020.d) Search engine for linked dataset. Retrieved from: <https://data.labs.kadaster.nl/>.
- Khodadadi, F., Dastjerdi, A. V., & Buyya, R. (2015, April). Simurgh: A framework for effective discovery, programming, and integration of services exposed in IoT. In *2015 International Conference on Recent Advances in Internet of Things (RIoT)* (pp. 1-6). IEEE.
- Kraak, M. J., & Ormeling, F. (2013). *Cartography: visualization of geospatial data*. CRC Press.
- Kolomiyets, O., & Moens, M. F. (2011). A survey on question answering technology from an information retrieval perspective. *Information Sciences*, 181(24), 5412-5434.
- Laparra, E., & Rigau, G. (2009, September). Integrating wordnet and framenet using a knowledge-based word sense disambiguation algorithm. In *Proceedings of the International Conference RANLP-2009* (pp. 208-213).
- Lafia, S., Turner, A., & Kuhn, W. (2018). *Improving discovery of open civic data*.
- Leseva, S., Stoyanova, I., & Todorova, M. (2018). Classifying Verbs in WordNet by Harnessing Semantic Resources. *Proceedings of CLIB*, 115-125.
- Lopez, V., Unger, C., Cimiano, P., & Motta, E. (2013). Evaluating question answering over linked data. *Journal of Web Semantics*, 21, 3-13.
- Li, G., Yan, L., & Ma, Z. (2019). Pattern match query over fuzzy RDF graph. *Knowledge-Based Systems*, 165, 460-473.
- Li, W., Zhou, X., & Wu, S. (2016). An integrated software framework to support semantic modeling and reasoning of spatiotemporal change of geographical objects: A use case of land use and land cover change study. *ISPRS International Journal of Geo-Information*, 5(10), 179.
- Lin, F., & Krizhanovsky, A. (2011). Multilingual ontology matching based on wiktionary data accessible via sparql endpoint. *arXiv preprint arXiv:1109.0732*.
- Lu, M., Sun, X., Wang, S., Lo, D., & Duan, Y. (2015, March). Query expansion via wordnet for effective code search. In *2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER)* (pp. 545-549). IEEE.
- Lutz, M., & Klien, E. (2006). Ontology-based retrieval of geographic information. *International Journal of Geographical Information Science*, 20(3), 233-260.
- Patwardhan, S. (2003). *Incorporating dictionary and corpus information into a context vector measure of semantic relatedness* (Doctoral dissertation, University of Minnesota, Duluth).
- Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004, May). WordNet: Similarity-Measuring the Relatedness of Concepts. In *AAAI* (Vol. 4, pp. 25-29).
- Perkins, J. (2014). *Python 3 text processing with NLTK 3 cookbook*. Packt Publishing Ltd.
- Pivert, O., & Smits, G. (2020). Fuzzy Extensions of Databases. In *Fuzzy Approaches for Soft Computing and Approximate Reasoning: Theories and Applications* (pp. 191-200). Springer, Cham.

- Postma, M., van Miltenburg, E., Segers, R., Schoen, A., & Vossen, P. (2016, January). Open dutch wordnet. In Proceedings of the Eight Global Wordnet Conference, Bucharest, Romania.
- Punjani, D., Singh, K., Both, A., Koubarakis, M., Angelidis, I., Bereta, K., ... & Lange, C. (2018, November). Template-based question answering over linked geospatial data. In Proceedings of the 12th Workshop on Geographic Information Retrieval (pp. 1-10).
- Rahmani, A. (2017, October). Adapting google translate for English-Persian cross-lingual information retrieval in medical domain. In 2017 Artificial Intelligence and Signal Processing Conference (AISP) (pp. 43-46). IEEE.
- Mai, G., Yan, B., Janowicz, K., & Zhu, R. (2019, June). Relaxing unanswerable geographic questions using a spatially explicit knowledge graph embedding model. In The Annual International Conference on Geographic Information Science (pp. 21-39). Springer, Cham
- Mandl, T. (2008). Recent developments in the evaluation of information retrieval systems: Moving towards diversity and practical relevance. *Informatica*, 32(1).
- Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 39-41.
- Mitra, B., Rosset, C., Hawking, D., Craswell, N., Diaz, F., & Yilmaz, E. (2019). Incorporating query term independence assumption for efficient retrieval and ranking using deep neural networks. arXiv preprint arXiv:1907.03693.
- Mur, J. (2008). Off-line answer extraction for question answering. University Library Groningen.
- O'Looney, J. (2000). Beyond maps: GIS and decision making in local government. ESRI, Inc.
- Sawant, U., Garg, S., Chakrabarti, S., & Ramakrishnan, G. (2019). Neural architecture for question answering using a knowledge graph and web corpus. *Information Retrieval Journal*, 22(3-4), 324-349.
- Sasaki, Y., & Fellow, R. (2007). The truth of the F-measure, Manchester: MIB-School of Computer Science. University of Manchester.
- Scheider, S., Nyamsuren, E., Krüger, H., & Xu, H. (2020). Geo-analytical question-answering with GIS. *International Journal of Digital Earth*, 1-14.
- Segev, A., & Gal, A. (2008). Enhancing portability with multilingual ontology-based knowledge management. *Decision Support Systems*, 45(3), 567-584.
- Sequeira, L. N., Moreschi, B., Cozman, F. G., & Fontes, B. (2020). An Empirical Accuracy Law for Sequential Machine Translation: the Case of Google Translate. arXiv preprint arXiv:2003.02817.
- Sun, T., Xia, H., Li, L., Shen, H., & Liu, Y. (2019). A Semantic Expansion Model for VGI Retrieval. *ISPRS International Journal of Geo-Information*, 8(12), 589.
- Tan, B., Shen, X., & Zhai, C. (2006, August). Mining long-term search history to improve search accuracy. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 718-723).
- Teufel, S. (2007). An overview of evaluation methods in TREC ad hoc information retrieval and TREC question answering. *Evaluation of text and speech systems*, 163-186.
- Tóth, K. (2012). A conceptual model for developing interoperability specifications in spatial data infrastructures. Office for Official Publications of the European Commission.
- Unger, C., Freitas, A., & Cimiano, P. (2014, September). An introduction to question answering over linked data. In Reasoning Web International Summer School (pp. 100-140). Springer, Cham.
- Usbeck, R., Ngomo, A. C. N., Conrads, F., Röder, M., & Napolitano, G. (2018). 8th Challenge on Question Answering over Linked Data (QALD-8). *language*, 7, 1.
- Vechtomova, O., Robertson, S., & Jones, S. (2003). Query expansion with long-span collocates. *Information Retrieval*, 6(2), 251-273.
- Wieleman, J. G. (2019). The semantic structure of spatial questions in human geography (Master's thesis).
- Xu, H., Hamzei, E., Nyamsuren, E., Krüger, H., Winter, S., Tomko, M., & Scheider, S. (2020). Extracting interrogative intents and concepts from geo-analytic questions. *AGILE: GIScience Series*, 1, 1-21.
- Zuva, K., & Zuva, T. (2012). Evaluation of information retrieval systems. *International journal of computer science & information technology*, 4(3), 35.