Improving the Performance of Arabic Information Retrieval Systems: The Issue of Resolving Word Sense Disambiguation

Wafya Hamouda[1], Abdulfattah Omar[2, 3], Yasser Muhammad Naguib Sabtan[4,5], &Waheed M. A. Altohami[2,6]

[1] Department of Foreign Languages, Faculty of Education, Tanta University, Tanta, Egypt

[2] Department of English, College of Science & Humanities, Prince Sattam Bin Abdulaziz University, Alkharj 11942, Saudi Arabia

[3] Department of English, Faculty of Arts, Port Said University, Port Said, Egypt

[4] Department of English Language and Literature, College of Arts and Applied Sciences, Dhofar University, Oman

[5] Department of English, Faculty of Languages and Translation, Al-Azhar University, Cairo, Egypt

[6] Department of Foreign Languages, Faculty of Education, Mansoura University, Egypt

Correspondence: Dr. Waheed M. A. Altohami, Associate Professor of Linguistics, Department of English, College of Science & Humanities, Prince Sattam Bin Abdulaziz University, Alkharj 11942, Saudi Arabia. E-mail: w.altohami@psau.edu.sa

Received: October 4, 2023 Accepted: November 13, 2023 Online Published: November 24, 2023

doi:10.5430/wjel.v14n1p297 URL: https://doi.org/10.5430/wjel.v14n1p297

Abstract

This study aimed at assessing the performance and efficacy of the retrieval information (IR) systems implemented in three widely used search engines (Google, Bing, and Yahoo), specifically with regard to the challenge of word sense disambiguation in Arabic texts. Such a challenge has been confirmed to negatively influence the retrieval of the most relevant documents. Therefore, we extended the paradigm of using computational methods and natural language processing (NLP) tools, primarily tailored for processing English texts, to explore morphosyntactic as well as lexical issues disturbing the accuracy of Arabic IR systems. Findings revealed striking disparities in the efficacy of IR systems integrated into these search engines, which can be attributed to four principal challenges: (a) the intricate morpho-syntactic structures inherent in Arabic; (b) the idiosyncratic orthographical system of the Arabic script; (c) the multifaceted semantic flexibility of certain lexical elements; and (d) the intriguing diaglossic nature of Arabic, allowing for the coexistence of multiple linguistic varieties within a single discourse situation. Drawing from these findings, a series of solutions rooted in supervised machine learning techniques, including clustering models and adaptations based on geographic locations, are proposed. Moreover, the study advocates for the capacity of search engines to interpret queries across all Arabic varieties, encompassing vernacular dialects. Furthermore, the importance of search engines accommodating queries irrespective of the specific language adopted by users is underscored. While the research primarily centers on Arabic, its implications resonate beyond this language alone. By applying computational methodologies originally designed for English to Arabic, the study not only addresses the challenges specific to Arabic IR systems but also contributes valuable insights that transcend linguistic boundaries. Through a comparative lens, issues like word sense disambiguation between Arabic and English are juxtaposed, extracting lessons that can inform advancements in information retrieval for both languages.

Keywords: ambiguity, Arabic, computational linguistics, information retrieval, NLP, word sense disambiguation

1. Introduction

The evolution of information retrieval systems has been remarkable over the past few decades, with advancements ranging from basic keyword searches to complex natural language processing (NLP) and machine learning techniques (Gao et al., 2023). The ultimate objective of these systems has remained consistent, which is to provide users with the required information in a timely and precise manner (George & George, 2023). Bibliographic databases were among the first information retrieval systems, enabling users to search for articles and publications through keywords or subject headings. Initially implemented in libraries and academic institutions, these databases soon became ubiquitous across various industries with the increasing availability of online information (Rubin, 2017). As the volume of available information increased, the retrieval systems used to locate it also became more intricate.

During the early stages of the internet, search engines such as AltaVista and Yahoo predominantly utilized keyword matching as the primary approach to providing search outcomes (Almukhtar, Mahmoodd, & Kareem 2021). However, these systems were easily manipulated by keyword stuffing and other tactics that made it difficult to find relevant information (Duka, Sikora, & Strzelecki, 2023). In response, search engines began incorporating more advanced algorithms and techniques to better understand the user's intent behind each search query (Nagpal & Petersen, 2021). More recently, natural language processing and machine learning have become increasingly important in information retrieval (IR). These techniques allow search engines to understand not just the keywords used in a search query but also the context in which those keywords are used (Lauriola, Lavelli, & Aiolli, 2022).

The accuracy of IR systems and search engines, therefore, depends on their ability to disambiguate words with multiple meanings, as the

incorrect sense can result in irrelevant search results (Omar & Aldawsari, 2020). The accuracy of an IR system is a measure of how well it retrieves relevant documents in response to a user's query (Husain, 2020; Hersh, 2020). One of the most significant challenges in achieving high accuracy on IR, however, is the problem of word sense ambiguity (Elayeb, 2019). Word sense ambiguity, also referred to as lexical ambiguity, is the phenomenon where a word has multiple meanings, and it can be challenging for an IR system to determine which meaning is intended in a given context (Agirre & Edmonds, 2007). When a user enters a query that contains an ambiguous word, an IR system must disambiguate the word to determine the intended meaning. If the system fails to disambiguate the word correctly, it may retrieve documents that are irrelevant to the user's needs, resulting in low accuracy. For example, if a user searches for 'apple,' an IR system may retrieve documents related to the fruit, the technology company, or the record label. If the system fails to disambiguate the term correctly, it may retrieve documents that are not relevant to the user's needs, leading to a lower accuracy rate.

Despite the improvements achieved, lexical ambiguity continues to pose a significant challenge that affects the quality and dependability of these systems. In IR applications, lexical ambiguity can lead to reduced precision and recall (Shaalan, Hassanein, & Tolba, 2017). When a query contains an ambiguous term, the retrieval system may miss some relevant documents that contain the intended meaning of the term, reducing the recall of the system. At the same time, the retrieval system may miss some relevant documents that contain the intended meaning of the term, reducing the recall of the system (Jain et al., 2022). When a query contains an ambiguous term, the retrieval system needs to analyze the context of the term to determine its intended meaning. This requires additional computational resources and time, leading to increased complexity and slower response times. This can have a significant impact on the overall performance of the system, especially in cases where a large number of queries need to be processed in real-time (Kejriwal, Knoblock, & Szekely, 2021). These challenges undoubtedly have adverse impacts on user satisfaction. Users expect information retrieval systems to return relevant and accurate results in response to their queries. When a system returns irrelevant or inaccurate results due to lexical ambiguity, users may become frustrated and dissatisfied with the system. This can lead to a decrease in user engagement and adoption, impacting the overall success of the system (Kumar & Santhosh, 2020).

The problem of lexical ambiguity is not limited to any particular language; it is universal, as in almost all languages, many words have multiple meanings (Arbaaeen, & Shah, 2020). Hence, incorporating methods that can help disambiguate words and accurately identify their intended senses in context is crucial to ensuring the precision and consistency of IR systems. To tackle the problem of lexical ambiguity in information retrieval (IR) systems, various techniques have been developed. These techniques can best be classified under the umbrella of word sense disambiguation (WSD).

WSD refers to identifying the intended meaning of a word that has multiple meanings based on the context in which it is used (Sheu et al., 2011). WSD is an NLP task that requires identifying the intended meaning of a word within a particular context (Zilli et al., 2008). This involves selecting the accurate sense of a word from a range of possible senses by considering factors such as nearby words, the word's part of speech, and other contextual information. In the context of IR systems and other NLP applications, WSD plays a critical role in improving the accuracy of these systems. In IR, WSD helps to ensure that the retrieved documents are relevant to the user's query by disambiguating the meaning of the query terms (Rahman & Borah, 2022). WSD is particularly useful in cases where the query terms have multiple meanings, as it can help to avoid retrieving irrelevant documents that contain the wrong sense of the query term (Scarlini, Pasini, & Navigli, 2020).

WSD is a challenging task in NLP, primarily due to the complexity of natural language and the fact that the same word can have multiple meanings depending on the context (Abderrahim & Abderrahim, 2022). As a result, researchers have developed various approaches to WSD, including supervised and unsupervised learning methods, rule-based systems, and hybrid approaches that combine multiple techniques. The process of WSD involves using computational methods to automatically disambiguate words with multiple meanings in a given context. This is done by considering the surrounding words, the syntactic structure of the sentence, and any other relevant contextual features. Once the correct sense of a word has been determined, IR systems can then retrieve only those documents that are relevant to the user's query, resulting in more accurate search results (Kaddoura, S., Ahmed, R. D., & Hemanth, 2022). Despite the development of various WSD techniques, there are inherent limitations to these methods, particularly in languages like Arabic, which have unique linguistic features that are not typically considered in standard IR systems.

Despite the development of multiple techniques for WSD to tackle linguistic ambiguity in NLP and IR systems, they still encounter restrictions when applied to Arabic (Mesfar, 2010). This is due to the specific linguistic features of Arabic, which are distinct from those of European languages in terms of phonetics, morphology, syntax, and semantics. As a Semitic language, Arabic poses distinctive obstacles for researchers and developers of NLP applications intended for Arabic speech and text (Omar, Elghayesh, & Kassem, 2019). The writing system of Arabic, which is entirely different from Western languages, can have an impact on linguistic ambiguity in natural language processing (NLP) applications, including information retrieval, due to certain characteristics of the language. Consequently, the success of IR systems in Arabic depends on their ability to account for these unique linguistic features (Elayeb, 2019). Failure to do so may result in inaccurate disambiguation, leading to the retrieval of irrelevant documents or missing relevant ones (Shaalan et al., 2018).

Previous WSD methods for Arabic often relied on simplistic approaches that treated each word as an independent unit, neglecting the rich morphological information available in the language. These methods failed to capture the nuances of word senses that can be distinguished based on morphological variations. Arabic is a highly inflectional language, which means that words undergo various morphological changes to indicate grammatical features such as tense, gender, number, and case. These morphological changes often result in different word forms or stems, each representing a different sense or meaning of the word. Therefore, considering the

morphological properties of Arabic is crucial for accurate word sense disambiguation.

In light of this argument, this paper proposes the integration of morphological analyzers and morphological disambiguation tools into WSD models to extract the possible stems or word forms of a given word. By considering the different stems and their associated senses, these models can better disambiguate word senses in Arabic. The premise is that integrating the morphological properties of Arabic is essential for accurate word sense disambiguation. By considering the various stems, morphological variations, and lemmas associated with a word, WSD models can better capture the different senses and improve the disambiguation process in Arabic NLP tasks.

Therefore, the present study examines whether the integration of morphological analyzers and morphological disambiguation tools can effectively address writing ambiguity in Arabic information retrieval applications within WSD models. To achieve this main objective, it seeks to address three research questions:

1. How does the integration of morphological analyzers and morphological disambiguation tools impact the disambiguation accuracy and overall performance of WSD models in resolving writing ambiguity in Arabic information retrieval applications?
2. Can the integration of morphological information enhance the performance of WSD models in low-resource settings for Arabic information retrieval?
3. How does the performance of WSD models with morphological integration compare to other existing techniques for addressing writing ambiguity in Arabic information retrieval?

The remainder of this study is organized as follows: Section 2 provides a brief survey of the methods used to handle Arabic writing ambiguity. This section presents an overview of the existing approaches and techniques that have been employed to address the challenges posed by the Arabic writing system in information retrieval. In Section 3, the paper presents the methods and procedures used in the study. This section outlines the research methodology, including data collection, experimental design, and evaluation metrics. The steps taken to analyze the impact of the Arabic writing system on information retrieval accuracy are described in detail. Section 4 reports the findings and results of the study and analyzes the effects of the Arabic writing system on the accuracy of information retrieval systems. It examines how the absence of explicit word boundaries and diacritical marks in Arabic script can introduce ambiguity and impact retrieval performance. The findings shed light on the challenges posed by the writing system and provide insights into potential solutions. Finally, in Section 5, the paper concludes by summarizing the main points discussed throughout the study. It highlights the implications of the findings for future research in Arabic information retrieval and suggests practical recommendations to enhance the accuracy of retrieval systems. The conclusion section emphasizes the importance of addressing the impact of the writing system on information retrieval and underscores the need for further investigations and advancements in this area.

2. Literature Review

In recent years, significant research has been dedicated to addressing the issue of word sense disambiguation in Arabic information retrieval. One of the most significant factors contributing to linguistic ambiguity and negatively impacting the accuracy of information retrieval systems is the Arabic writing system. Saadi and Belhadef (2020) indicated that the Arabic writing system plays a crucial role in causing linguistic ambiguity and adversely affecting the precision of information retrieval systems. Arabic is written in a connected script, with letters joined together within a word. This unique feature presents difficulties for tasks such as optical character recognition, text segmentation, and word boundary detection in natural language processing (NLP) applications. Incorrect segmentation or misinterpretation of ligatures can thus contribute to errors in lexical analysis and exacerbate lexical ambiguity.

Almanaseer et al. (2021) agree that the Arabic writing system is a major contributor to linguistic ambiguity and has a detrimental effect on the precision of information retrieval systems. They explain that Arabic script lacks explicit markers for word boundaries in a sentence, leading to different interpretations and potential ambiguity in information retrieval. Additionally, Arabic script lacks diacritical marks, specifically short vowels, which are crucial for disambiguating words. Without these diacritics, multiple interpretations become possible, making it challenging for information retrieval systems to accurately match queries with relevant documents. However, it is important to recognize that Arabic writing ambiguity is not an isolated issue. It is intricately linked with other factors that further complicate the problem. These factors include the unique morphological and syntactic system in Arabic, dialectical variation, and diglossia.

Al-Zoghby et al. (2013) have demonstrated that the Arabic writing system is characterized by a root-based morphology, in which words are formed from a consonantal root. The positioning of vowels and additional consonants around the root can give rise to multiple words that share the same root but possess different meanings. Omar and Hamouda (2020) further explain that Arabic words are derived from root letters and can undergo various modifications, such as the addition of prefixes or suffixes and internal vowel changes. These modifications can significantly alter the meaning of words or introduce ambiguity. Consequently, the intricate morphological structure of Arabic poses significant challenges for information retrieval systems, necessitating the utilization of sophisticated algorithms and resources to accurately analyze and disambiguate Arabic words.

Alnaied et al. (2020) provide additional insights by highlighting that the Arabic language boasts a diverse vocabulary, with numerous word forms derived from root letters through morphological processes. This linguistic richness gives rise to multiple variations of words, involving the addition of prefixes, suffixes, and changes in internal vowels. Consequently, a single query term can manifest in multiple forms, each potentially associated with distinct meanings. This variability poses further challenges to the disambiguation of word senses in information retrieval applications.

In order to tackle these challenges, different methods have been used for handling Arabic writing ambiguity in information retrieval, including linguistic approaches, word sense disambiguation techniques, and machine learning-based methods. Most linguistic approaches have been devised to harness linguistic knowledge and resources in order to disambiguate words and address interpretation challenges in Arabic text. These approaches make use of various linguistic theories, including semantic, psycholinguistic, pragmatic, and discourse theories. By incorporating these theoretical frameworks, researchers aim to enhance the accuracy and precision of Arabic text analysis and interpretation. According to Hammo (2008), semantic analysis is among the earliest approaches aimed at tackling linguistic ambiguities in Arabic information retrieval. Hammo (2008) argues that semantic analysis offers an effective method for exploring the meaning of words and phrases within a wider context, thus aiding in the resolution of ambiguities. To this end, various semantic theories have been developed to elucidate the essence of linguistic ambiguity and capture generalizations pertaining to the context-dependent nature of word meaning.

The underlying principle behind semantic approaches to address the issue of ambiguity is the recognition that writing ambiguity is primarily associated with word meaning rather than structural aspects, which can lead to multiple interpretations. Semantic studies have proposed different approaches to determining the correct sense in ambiguous sentences. Among the most prominent approaches are semantic relatedness or interconnections, cognitive topology, and lexical networks. These approaches rely on external resources, such as lexical databases, ontologies, and thesauri, to disambiguate word meanings. They use the knowledge contained in these resources to identify the correct sense of an ambiguous word (Rodd, 2018). Although these approaches can be usefully used in disambiguating words in different contexts and thus helping in the development of reliable NLP applications, they tend to perform better when there is a limited amount of data, but they require high-quality lexical resources, which may not always be available.

Psycholinguistic studies have explored the mental processes involved in resolving lexical ambiguity and how the human brain responds to such ambiguity. Chomsky's concept of linguistic competence has been used as the basis for traditional psycholinguistic approaches to lexical ambiguity, which aim to understand the knowledge and abilities required for language comprehension and production (Chomsky, 1965). These studies have investigated the universality of the problem of lexical ambiguity across different languages and cultures. Traditional approaches have generally viewed lexical ambiguity as a problem to be avoided because it can lead to confusion and misunderstanding. However, recent studies in psycholinguistics have challenged this view and suggested that ambiguity can be beneficial as it allows for words to be used in different contexts, facilitating language processing. Such studies have explored the potential advantages of lexical ambiguity and the role it plays in the interpretation of language.

The premise of recent psycholinguistic approaches to lexical ambiguity is to study how humans process language and how they disambiguate words with multiple meanings. These approaches aim to understand how the human mind makes sense of ambiguous words and to use this knowledge to improve natural language processing systems. These approaches are based on the idea that humans use various cues, such as syntax, semantics, and pragmatics, to disambiguate words in context (Navigli, 2012). These cues help humans identify the intended meaning of a word based on the surrounding words, the sentence structure, and the overall discourse. Recent psycholinguistic approaches often use experimental methods, such as eye-tracking and reaction time studies, to investigate how humans process ambiguous words. These studies aim to identify the cognitive processes involved in word sense disambiguation and to determine which cues humans rely on the most.

Discourse-based approaches have shown the potential to resolve lexical ambiguity by integrating context and discourse into the analysis. These approaches have been further explored in corpus-based studies that use electronic corpora, demonstrating the effectiveness of using discourse to disambiguate meaning. They aim to resolve the issue of lexical ambiguity by considering the context in which a word is used. These approaches take into account the surrounding words and phrases, as well as the broader discourse, to determine the most appropriate meaning of the ambiguous word (Arbaaeen & Shah, 2020). One common discourse-based approach is known as the 'one sense per discourse' hypothesis, which suggests that a word is likely to have only one sense within a given discourse context. This approach relies on the assumption that the overall meaning and purpose of the discourse will help to disambiguate the meaning of individual words (Arbaaeen & Shah, 2020). Another approach is known as 'lexical cohesion,' which refers to the way in which words within a text are related to one another. By analyzing the patterns of lexical cohesion, it is possible to identify the most likely meaning of an ambiguous word based on the words and phrases that surround it.

Other discourse-based approaches include 'pragmatic inference,' which involves using knowledge about the speaker's intentions and beliefs to disambiguate an ambiguous word, and "semantic frames," which involve using knowledge about the typical contexts in which a word is used to disambiguate its meaning. While linguistic approaches offer valuable insights, they can be limited by the complexities and variations of the Arabic language.

Word sense disambiguation (WSD) techniques aim to identify the correct meaning of a word based on its context. Various WSD methods have been adapted to handle Arabic writing ambiguity. Knowledge-based approaches rely on dictionaries, lexical resources, and ontologies to assign the most appropriate sense to a word. Corpus-based approaches utilize large text corpora to identify word usage patterns and disambiguate based on statistical analysis. Hybrid approaches combine knowledge-based and corpus-based techniques to enhance disambiguation accuracy (Faizullah et al., 2023). These WSD techniques provide effective means to address writing ambiguity in Arabic information retrieval systems. WSD techniques have been developed with the aim of improving the performance of retrieval systems, as correct disambiguation of words can lead to more accurate retrieval of relevant documents for user queries. This highlights the importance of addressing lexical ambiguity in the development of IR systems, as it can significantly impact the effectiveness of the

system in providing relevant information to users.

Another approach to WSD is ontology-based, which relies on ontologies to represent knowledge about a particular domain. An ontology is a formal representation of concepts and their relationships within a given domain. This approach involves mapping the words in a text to the concepts in the ontology and then disambiguating the senses of the words based on their relationships to other concepts in the ontology (Agirre & Edmonds, 2007). This approach has been used in various fields, such as medical informatics and bioinformatics, where ontologies are used to represent the complex relationships between medical terms and biological concepts (Senanayake et al., 2023). However, one of the challenges of the ontology-based approach is the difficulty in creating and maintaining the ontology, as it requires domain expertise and extensive manual effort (Agirre & Edmonds, 2007).

The knowledge-based approach to WSD involves the use of machine-readable dictionaries and semantic lexicons that contain information about word senses and their relationships to other words. This approach is based on the idea that knowledge about a word can be used to disambiguate its sense in a particular context (Navigli, 2012). This approach has been shown to be effective in WSD tasks, especially when used in combination with other methods such as corpus-based techniques. However, one of the limitations of this approach is the need for extensive knowledge resources, which may not always be available or may be difficult to acquire in some domains. Overall, while each approach to WSD has its strengths and limitations, recent research has shown that combining different methods can improve the performance of WSD systems (Nguyen et al., 2018). Moreover, advances in machine learning techniques, such as neural networks and deep learning, have shown promising results in WSD tasks, providing a new direction for future research in this area (Huang et al., 2019).

Ontology-based techniques have been developed as a way to overcome the limitations of both dictionary-based and knowledge-based approaches to WSD. Ontologies are a popular method in IR systems and are designed to enable the resolution of lexical ambiguity by using a network of semantic concepts to draw inferences (Agirre & Edmonds, 2007). The idea behind ontology-based techniques is that searches in IT should be based on meaning and inference rather than just on literal strings. By using ontologies, IR systems can understand the relationships between search items and concepts, which can lead to more accurate and relevant results. However, ontology-based techniques can be challenging to implement because they require a deep level of conceptualization, and complex query languages may be necessary. Despite these challenges, ontology-based techniques are a promising way to improve WSD and enhance the performance of IR systems (Zilli et al., 2008).

Machine Learning-Based Methods: Machine learning algorithms have gained popularity for handling Arabic writing ambiguity due to their ability to automatically learn patterns from data. As far as WSD is concerned, machine learning-based methods are classifiable into four categories. The first group is based on probabilistic models that estimate a set of probabilistic parameters, which signify the conditional probability of each category in a specific context, by means of specific algorithms (e.g., the Naive Bayes algorithm) to estimate other new parameters. These methods work competitively in WSD (cf. Agirre & Edmonds, 2007). The second group is based on the similarity of the examples, as it takes into account a similarity metric to perform the WSD process. One example representing such methods is the Vector Space Model (VSM), which compares new examples to a set of prototypes and assigns the sense of the most similar prototype. The third group is based on discriminating rules that are associated with the sense of each word. That is, to determine one sense of the senses of a polysemous word for example, the IR system selects the sense that verifies some such rules (Navigli, 2012).

The last group of machine learning-based methods is based on decision lists, which are ordered lists of rules of the form considered as weighted ‘if-then-else’ rules. High weights come at the top of the decision tree and low weights at the bottom. Yet, decision trees are not frequently used in WSD. Support vector machines, decision trees, and neural networks are commonly used in this context. These algorithms are trained on annotated Arabic datasets, where words are disambiguated based on their surrounding context. By learning from these examples, machine learning models can generalize and accurately disambiguate words in real-world Arabic text. However, the success of machine learning-based methods relies on the availability of high-quality training data and the careful design of features and representations.

3. Methodology

This section outlines the rationale beyond the selection of particular search engines for comparing their IR performance as far as the issue of lexical ambiguity is concerned, sketches the final dataset, and explains the procedure of data analysis. Given the scope of the current research, which seeks to explore the challenge of lexical ambiguity that hinders developing effective IR systems for Arabic, we selected three popular search engines to be compared with reference to their IR performance: Google, Bing, and Yahoo. Chris (2019) reported that these three search engines were among the most widely used search engines, according to the reports of the Digital Marketing Agency. For the sake of a valid and reliable comparison of the three search engines, a set of criteria was applied based on previous research (cf. White, 2016; Lupu, Mayer, Kando, & Trippe, 2017). These criteria are (a) retrieving relevant documents and excluding irrelevant ones, (b) similar response time providing that all documents under exploration contribute similarly to the final dataset, (c) the mechanism of IR retrieval, and (d) the techniques implemented for enhancing IR.

In order to explore the precision ratio of the IR system performance in each search engine, a particular query is used, and the precision query is calculated by the following equation: number of only retrieved documents/total number of retrieved documents x 100. Figure (1) shows the process of evaluating IR systems.

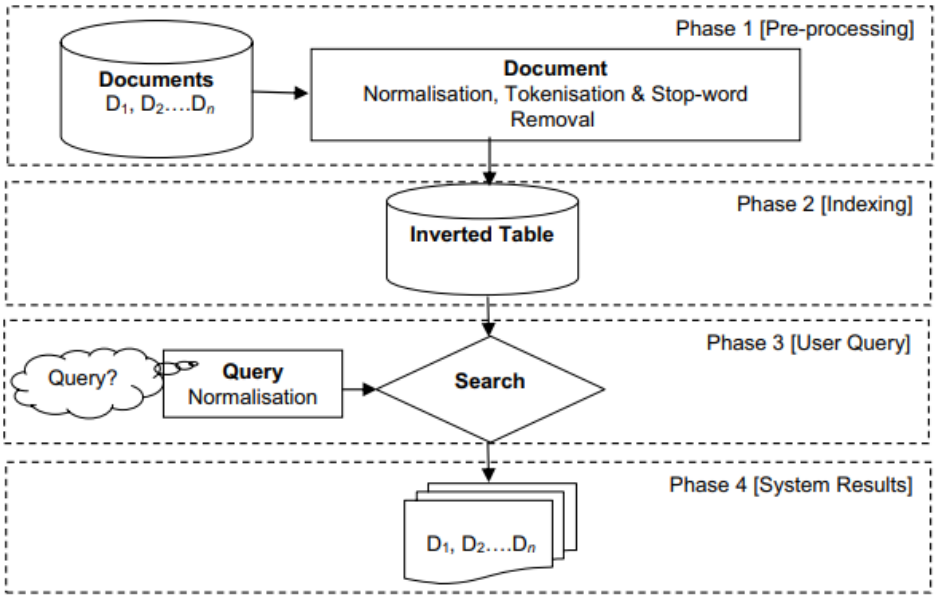

Figure 1. Phases of standard IR (Atwan et al., 2016, p. 247)

In the pre-processing phase, the output documents are normalized by removing redundant words to create a more cohesive body of documents and tokenized by removing sensitive data and replacing it with specific values. Then, stop-words, i.e., words used across all documents, are removed. In the phase of indexing, an inverted, unordered table is ordered to improve the efficiency of the target query during searching by offering faster results. Such evaluation measures are paramount to assessing the efficacy of the IR system to make well-informed decisions to meet the information needs of users. Assuming that the selected search engines differ with regard to their recall performance, we used four main queries representing the most popular issues in the Arab countries in 2023. Table (1) shows the list of queries used for retrieving relevant documents and the total number of retrieved sites.

Table 1. List of queries used for retrieving relevant documents from Google, Bing and Yahoo

No.	Query	Translation	Total of retrieved sites		
			Google	Bing	Yahoo
1	زلزال المغرب	Morrocco Earthquake	33,100,000	123,000	1,980,000
2	موسم الرياض	Riyadh Season	22,300,000	244,000	3,880,000
3	أوكرانيا	Ukraine	13,700,000	282,000	2,010,000
4	عاصفة دانيال	Storm Daniel	11,400,000	224,000	226,000
5	دوري روشن	Roshn League	9,060,000	183,000	1,900,000

As shown in Table (1), there are four multi-word queries (i.e., عاصفة دانيال, زلزال المغرب, موسم الرياض, and دوري روشن) and one-word query (*i.e.*, أوكرانيا Ukraine). Based on the results of these queries, relevant and irrelevant documents are identified. Finally, the reasons for lexical ambiguity are outlined and discussed.

4. Findings and Discussion

By comparing the number of relevant documents produced by Google, Bing, and Yahoo search engines based on the five queries, we found that Google ranked the first as far as the multi-word queries are concerned. While Yahoo ranked first with regard to the one-word query. For further statistics, see Table (2).

Table 2. Query-based total of relevant retrieved sites

No.	Query		Translation	Total of relevant retrieved sites		
				Google	Bing	Yahoo
1	Multi-word queries	زلزال المغرب	Morrocco Earthquake	15,888,000 [48%]	51,660 [45%]	930,600 [47%]
2		موسم الرياض	Riyadh Season	10,258,000 [46%]	95,160 [39%]	1,590,800 [41%]
3		عاصفة دانيال	Ukraine	5,700,000 [50%]	91,840 [41%]	99,440 [44%]
4		دوري روشن	Storm Daniel	3,895,800 [43%]	71,370 [37%]	741,000 [39%]
5	One-word query	أوكرانيا	Roshn League	6,028,000 [44%]	169,200 [60%]	1,346,700 [67%]

The data analysis showed that the performance of the three IR systems represented by the three search engines is negatively influenced by lexical ambiguity. Lexical ambiguity was the first reason for excluding the largest number of retrieved sites, and it was ascribed to the following reasons:

4.1 The Morpho-Syntactic Nature of Arabic

As a Semitic language, Arabic has a highly rich derivational morphological system, as any word is built from a root consisting of three letters, representing an index in an IR system. For instance, the lexical items 'حرب', 'حروب', 'محارب', and 'محارب' are derived from the three-letter discontinuous consonantal root and stem 'حرب'. Therefore, stemming, i.e., reducing the form of a lexical item by removing all suffixes, prefixes, and diacritics (*i.e.*, special characters used above and below Arabic letters, *e.g.*, damma, fatha, kasrah, and hamza), becomes really challenging in NLP applications. In any IR system, stemming is crucial to match them with other synonyms across the retrieved documents. The complexities of Arabic morphology due to non-concatenating affixes and clitics within Arabic words also pose a real challenge in the process of tokenization (Soudi et al. 2007). That is, one word might have up to four or more tokens. For instance, based on the discontinuous consonantal root 'سأل', tokens such as 'سؤال', 'مسؤول', 'مسئلة', 'مسؤولة', 'مسائل', and 'مسؤولية'. In other words, Arabic is agglutinative, *i.e.*, affixes could be added to a root to form new lexical items that belong to the same lexical category or different categories, each with a distinct lexical feature.

Furthermore, though a sentence could be nominal or verbal with a subject, a verb, and a complete thought (verb-subject-object, subject-verb-object, and object-verb-subject), a constituent might be deleted for a particular stylistic effect. For instance, Arabic allows the deletion of subject pronouns (the pro-drop feature) as long as they are recoverable based on the context. The result would be subjectless sentences (Chomsky, 1965), which are structurally ambiguous. Also, some structures are syntactically ambiguous due to syntactic movements and the absence of capitalization that shows where the sentence starts and where it ends. Further to this, structural ambiguity represents another challenge to IR systems. That is, the same stricture could be interpreted differently, *i.e.*, two or more meanings are to be accepted. For instance, 'رئيس الفريق السعودي' could mean either 'the chair of the Saudi team' or 'the Saudi chair of the team'. In the indexing phase of the IR systems evaluation, there are two noun phrases (*viz.*, the Saudi chair and the Saudi team) that trigger the two possible interpretations. Such unique and complex morhosyntactic nature of Arabic represents real challenges to NLP applications, including automatic text summarization methods (cf. Omar et al., 2023) and authorship detection (cf. Omar, 2021).

4.2 The Orthographical System of Arabic

Arabic is a phonetic language, but there is no one-to-one correspondence between letters and sounds. The shape of Arabic letters is rule-governed as it changes based on their position at the beginning, in the middle, or at the end of a word, *e.g.*, 'خ' is used in initial/final position; 'ـخـ' is used in medial position 'ـخ'; and 'ـ' is used in final position. Figure (2) displays an extract based on the query ' موسم الرياض', and notice the way the letter 'ف' is spelt initially, medially, and finally.

وعندما يحدث ذلك، فإن الصخور المحيطة بالصدع يحدث لها تشوه، فيتجه بعضها للأعلى أو للأسفل، مما يعني أن سطح الأرض يتحرك أيضا للأعلى أو للأسفل، وفي الصدوع العكسية، كالتي تسببت في زلزال المغرب الأخير، تتحرك الكتلة العلوية فوق مستوى الصدع، لمستوى أعلى من الكتلة السفلية، وهذا النوع من التصدع شائع في مناطق الانضغاط، مثل المناطق التي يتم فيها انزلاق إحدى الصفائح تحت أخرى، كما هو الحال في المغرب، حيث تنزلق الصفيحة الأفريقية تحت الأوروآسيوية.

Figure 2. A screenshot showing lack of clear sentence boundaries in Arabic texts

Also, punctuation marks have no strict rules as they are manipulated to perform different stylistic functions, which are crucial to understanding the meaning of a text (Dahl, 2018). Sometimes the absence of punctuation marks leads to the formation of one-sentence paragraphs. One more reason for lengthy sentences is the excessive use of connectors such as relative pronouns, conjunction letters, etc. Figure (3) displays an extract from https://www.aljazeera.net/ showing a lack of clear sentence boundaries in Arabic.

وتضمنت المسيرة في انطلاقتها مشاركة الفنان العالمي بيتبول بحفل افتتاح «موسم الرياض»، وإحياء فعاليات الليلة الأولى من الموسم، الأمر الذي سيرتقي بالحفل من الاهتمامات الإقليمية إلى العالمية، ويضاعف الإقبال الجماهيري على الفعاليات الأولية للموسم.

Figure 3. A screenshot showing lack of clear sentence boundaries in Arabic texts

Similarly, diacritics are problematic since they change the lexical meaning of a word. For instance, the lexical items 'قَدَم', 'قِدَم', 'قَدَمَ', and 'قُدُم'—which are homographs—mean 'foot' (noun), 'oldness' (noun), 'introduced' (verb), and 'forward' (adverb), respectively. Most writers nowadays do not care for these diacritics, thereby causing lexical ambiguity in many cases. Further to this, foreign and Arabized lexicons are used with different spellings, *i.e.*, there is no standard form for transliterated and calqued words. The documents under exploration showed

different spellings and transliterations for English words. For instance, 'الديموغرافية'(demography) is also spelt as 'الديموجرافية' and 'الديمغرافية'. These facts pose a real challenge when (pre)processing Arabic texts automatically via the IR systems built into the selected search engines. Equally important, the process of normalizing Arabic script is thus another big challenge that affects the performance of such IR systems. Even if such IR systems managed to normalize the Arabic script with all its complexities and render it easily recognized, diverse cases of lexical ambiguity still exist.

4.3 The Semantic Versatility of Some Lexical Items

Put simply, a lexical item is said to be semantically versatile if it denotes or connotes different senses in different contexts. Such semantic variation takes place if lexical items are polysemous, homonymous, or polyonomous, *i.e.*, they share various lexical semantic features. Indeed, Arabic is rich with lexemes whose meanings vary greatly, let alone the figurative use (or the sematic shift) of such lexemes. Such semantic complexity renders information retrieval from Arabic documents a difficult task due to issues related to lexical ambiguity (*cf*. Faizullah et al., 2023). The search engines in focus apparently lack a solid semantic-based searching system, which ensures more relevant information retrieval. For instance, the lexical item 'عاصفة' is used in some documents as a denotatum with the literal meaning of a violent weather condition of strong rain, wind, thunder, and lightning, as in ' وذكرت الأرصاد أن **العاصفة** اتجهت إلى الحدود المصرية، وستؤثر على أقصى غرب البلاد ولكن في هيئة منخفض متعمق' [Meteorology reported that the storm headed to the Egyptian border and would affect the far west of the country, but in the form of a deep depression]. Also, it was used figuratively as it was associated with strong feelings either positively or negatively, as in 'تصريح رسمي يثير **عاصفة** في الدوري السعودي . . .هل تغيرت ملامح البطل؟' [An official statement raises a storm in the Saudi League... Have the hero's features changed?]. Though such differences can be resolved by contextual clues, they still impede the pre-processing and normalization of Arabic data (Zait & Zarour, 2018).

As mentioned earlier, the standard Arabic script uses diacritics to distinguish the possible meanings of words. With the normalization of Arabic script, it becomes challenging to distinguish such meanings. For instance, handing the letter alif the same by the IR systems makes it difficult to distinguish the uses of the grammatical units 'إنَّ', 'أنْ', 'آن' and 'إنْ'. The word 'إنَّ' (translated as *that*) is used for emphasis, and it is followed by a nominal sentence as in ' قال مسؤولون أوكرانيون يوم الثلاثاء **إنَّ** روسيا قصفت بنية تحتية للموانئ وصوامع الحبوب بأوكرانيا في هجوم بطائرات مسيرة خلال الليل' [Ukrainian officials said on Tuesday that Russia bombed port infrastructure and grain silos in Ukraine in an overnight drone attack]. While 'أنْ' is followed by an infinitive as in ' إن كل من تعرض لكارثة طبيعية لا شك أنه شهيد، موجها العزاء **أن** يرحم الله ضحاياهم ومصر أجمعين؛ بسبب إعصار ليبيا للأشقاء في ليبيا' [Everyone who is exposed to a natural disaster is undoubtedly a martyr. We extend our condolences to the brothers in Libya. May God have mercy on their victims and all of Egypt because of the hurricane in Libya]. Yet, the word 'آن' is used as a time indicator, meaning 'it is time to' and it is also followed by an infinitive, as in ' وذكر أنه قد **آن** للمسئولين الرياضين أن يعملوا بجد في إنهاء التعاقدات المطلوبة قبل بدء الموسم الكروي الجديد' [He stated that it is time for sports officials to work hard to finalize the required contracts before the start of the new football season]. Finally, 'إنْ' is used to make a condition (=if), as in ' قد يحصل زلزلان قويان بنفس المنطقة في فترة قصيرة، و**إن** حدث ذلك فستكون النتائج كارثية' [Two strong earthquakes may occur in the same area in a short period, and if this happens, the results will be catastrophic].

This finding is consistent with the findings of previous studies, which emphasized that the senses of words are crucial to the retrieval of relevant or irrelevant documents (Krovetz & Croft, 1992; Abderrahim & Abderrahim, 2022; Saidi et al., 2022). Conducting a word sense analysis of irrelevant documents could help with disambiguating them, thus improving the precision of IR systems. In other words, the distribution of highly frequent and rare senses would support conducting a deeper semantic analysis of the documents by means of clustering techniques.

4.4 The Diaglossic Nature of Arabic

Ferguson (1959) defined 'diglossia' as the use of two varieties of the same language in distinct situations to communicate diverse functions in the same speech community. In most Arab speech communities, three basic varieties are used: (a) Classical Arabic (high variety), which is restricted to the Quran, Prophetic speeches, and classical poetry; (b) Modern Standard Arabic (MSA) (low variety), which is used in formal settings such as media and academic settings; and (c) colloquial Arabic, a less-sophisticated form of MSA, which is used with family, friends, and others in everyday interactions, and it differs regionally (AlSuwaiyan, 2018). On the morpho-syntactic level, Classical Arabic and MSA share many aspects; whereas on the lexical level, they are markedly different (Simpson, 2019). In a document, more than variety might be used, for instance, "قبل أيام أقر عقيلة صالح وبرلمانه الفائدة في البنوك (الربا) تحدي لأمر الله، هل إعصار دانيال جاء ردا على سكوت الناس ورضاهم بإحلال الربا . **فأذنوا بحرب من الله ورسوله**" [A few days ago, Aguila Saleh and his parliament approved interest in banks (usury) in defiance of God's command. Did Hurricane Daniel come in response to the silence of the people and their satisfaction with the introduction of usury? Then take notice of war from Allah and His Messenger]. In the previous instance, Classical Arabic is used by incorporating a part of a Quranic verse (فأذنوا بحرب من الله ورسوله). Also, the whole utterance is rendered in MSA. Mixing both varieties in one document represents a real challenge to NLP applications as it accounts for many instances of lexical ambiguity. The reason is that high and low varieties differ in terms of syntactic structures, word usage, and stylistic features, and therefore data processing becomes difficult and even impossible.

Equally important, the data collected comprised diverse Arabic dialects, including Moroccan Arabic, Libyan Arabic, Egyptian Arabic, and Saudi Arabic. Each dialect has its own lexicon, morphology, and syntax, which are not clearly sketched. The apparent ignorance of the peculiar linguistic aspects of the wide array of Arabic dialects in the search engines under exploration accounts for the notable difference in their performance in retrieving relevant sites. This finding agrees with Farghaly and Shaalan (2009) and Obeid et al. (2019), who empirically

stressed that most of the available and widely used research engines cater to MSA, and that the current IR systems are trained to handle the morpho-syntactic and semantic features of MSA. Let alone the fact that most of the users of search engines are youth who naturally use colloquial language (*cf.* Azmi & Aljafari, 2015).

Also, it must be acknowledged that there is no standard writing system for colloquial Arabic due to massive dialectical variations. However, the ever-increasing number of social media networks supporting Arabic could be used by NLP specialists to outline the key linguistic features of concurrent Arabic dialects (Harrat et al. 2019; Hicham et al., 2023; Matrane et al., 2023). Therefore, integrating the linguistic features of recurrent Arabic dialects into the IR systems will enhance their performance, and offer more relevant information to search engine users. Furthermore, the pragmatic force of the same utterance might differ across a set of situations. Unfortunately, most of the NLP stemmers and parsers built in the three search engines under exploration are primarily designed to handle MSA data. This fact is affirmed by previous studies, which highlighted the peculiar linguistic features of the low and high varieties of Arabic. Such features should be taken into account by IR engineers in order to develop viable and trustworthy solutions to the problem of lexical ambiguity (Hijjawi & Elsheikh, 2015; Mustafa & Suleman, 2015; Rodd, 2018).

5. Conclusion

The purpose of this study was to investigate the intersection of computational linguistics and natural language processing, with a particular focus on the interplay between Arabic and English within this multidisciplinary domain. The primary aim was to evaluate the transferability and applicability of computational methods, originally crafted for processing English texts, when applied to the realm of Arabic information retrieval (IR) systems. To achieve this goal, the study undertook a comprehensive examination of the challenges inherent in developing effective IR systems for Arabic, with a specific emphasis on the task of word sense disambiguation. Through a comparative analysis of the performance of three prominent search engines—Google, Bing, and Yahoo—in retrieving the most pertinent documents, the intricate nature of the Arabic language, which poses formidable hurdles for the creation of accurate IR systems, was considered. The results highlighted obvious disparities in the efficacy of the IR systems integrated into these search engines. These disparities can be attributed to four fundamental challenges: (a) the morpho-syntactic intricacies that characterize Arabic; (b) the idiosyncratic orthographical system of the Arabic script; (c) the multifaceted semantic adaptability of certain lexical elements; and (d) the intriguing diaglossic nature of Arabic, permitting the simultaneous use of multiple linguistic varieties within a single discourse context.

Building upon the findings of this study, a spectrum of supervised machine learning-based solutions has been proposed, encompassing clustering models and region-specific adaptations. Furthermore, the integration of morphological analyzers and morphological disambiguation tools into search engines is recommended to heighten the disambiguation accuracy and overall performance of Word Sense Disambiguation (WSD) models, thus facilitating the resolution of writing ambiguity in Arabic information retrieval applications. Moreover, the study advocates for search engines to possess the capability to interpret queries across all Arabic dialects, including colloquial varieties. Recognizing the linguistic intricacies inherent in these dialects, the importance of accommodating these variations is underscored in the pursuit of more effective Arabic information retrieval systems.

Authors contributions

All authors contributed equally to writing, editing, and proofreading the manuscript.

Funding

This study is supported via funding from Prince Sattam bin Abdulaziz University, project number (PSAU/2023/R/1445).

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics approval

The Publication Ethics Committee of the Sciedu Press.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.

References

Abderrahim, M. A., & Abderrahim, M. E.-A. (2022). Arabic word sense disambiguation for information retrieval. *Transactions on Asian and Low-Resource Language Information Processing, 21*(4), 1-19. https://doi.org/10.1145/3510451

Agirre, E., & Edmonds, P. (2007). *Word sense disambiguation: Algorithms and applications*. Springer Science & Business Media.

Almanaseer, W., Alshraideh, M., & Alkadi, O. (2021). A deep belief network classification approach for automatic Diacritization of Arabic text. *Applied Sciences, 11*(11), 5228. https://doi.org/10.3390/app11115228

Almukhtar, F., Mahmoodd, N., & Kareem, S. (2021). Search engine optimization: A review. *Applied Computer Science, 17*(1), 70-80. https://doi.org/10.23743/acs-2021-07

Alnaied, A., Elbendak, M., & Bulbul, A. (2020). An intelligent use of stemmer and morphology analysis for Arabic information retrieval. *Egyptian Informatics Journal, 21*(4), 209-217. https://doi.org/10.1016/j.eij.2020.02.004

AlSuwaiyan, L. A. (2018). Diglossia in the Arabic Language. *International Journal of Language and Linguistics, 5*(3), 228-238. https://doi.org/10.30845/ijll.v5n3p22

Al-Zoghby, A. M., Ahmed, A. S., & Hamza, T. T. (2013). Arabic Semantic Web applications – A survey. *Journal of Emerging Technologies in Web Intelligence, 5*(1). https://doi.org/10.4304/jetwi.5.1.52-69

Arbaaeen, A., & Shah, A. (2020). Natural language processing-based question answering techniques: A survey. *2020 IEEE 7th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*. https://doi.org/10.1109/icetas51660.2020.9484290

Azmi, A., & Aljafari, E. (2015). Modern information retrieval in Arabic—catering to standard and colloquial Arabic users. *Journal of Information Science, 41*(4), 506-517. https://doi.org/10.1177/0165551515585720

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge.

Chris, A. (2019). *Top 10 search engines in the world*. Retrieved from https://www.reliablesoft.net/top-10-search-engines-in-the-world/

Dahl, A. (2018). The graphic and grammatical structure of written texts. *Studia Neophilologica*, *90*(sup1), 24-36. https://doi.org/10.1080/00393274.2018.1531248

Duka, M., Sikora, M., & Strzelecki, A. (2023). From web catalogs to Google: A retrospective study of web search engines sustainable development. *Sustainability, 15*(8), 6768. https://doi.org/10.3390/su15086768

Elayeb, B. (2019). Arabic word sense disambiguation: A review. *Artificial Intelligence Review, 52*(4), 2475-2532. https://doi.org/10.1007/s10462-018-9622-6

Faizullah, S., Ayub, M. S., Hussain, S., & Khan, M. A. (2023). A survey of OCR in Arabic language: Applications, techniques, and challenges. *Applied Sciences, 13*(7), 45-84. https://doi.org/10.3390/app13074584

Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing. *ACM Transactions on Asian Language Information Processing, 8*(4), 1-22. https://doi.org/10.1145/1644879.1644881

Ferguson, C. (1959). Diglossia. *WORD, 15*(3), 325-340.

Gao, J., Xiong, C., Bennett, P., & Craswell, N. (2023). *Neural approaches to conversational information retrieval.* Springer Nature.

George, A. S., & George, A. H. (2023). A review of ChatGPT AI's impact on several business sectors. *Partners Universal International Innovation Journal, 1*(1), 9-23. https://doi.org/10.5281/zenodo.7644359

Hammo, B. H. (2008). Towards enhancing retrieval effectiveness of search engines for diacritisized Arabic documents. *Information Retrieval, 12*(3), 300-323. https://doi.org/10.1007/s10791-008-9081-9

Harrat, S., Meftouh, K., & Smaili, K. (2019). Machine translation for Arabic dialects (survey). *Information Processing & Management, 56*(2), 262-273. https://doi.org/10.1016/j.ipm.2017.08.003

Hersh, W. (2020). *Information retrieval: A biomedical and health perspective.* Springer International Publishing.

Hijjawi, M., & Elsheikh, Y. (2015). Arabic language challenges in text based conversational agents compared to the English language. *International Journal of Computer Science and Information Technology, 7*(3), 1-13. https://doi.org/10.5121/ijcsit.2015.7301

Hicham, N., Karim, S., & Habbat, N. (2023). Customer sentiment analysis for Arabic social media using a novel ensemble machine learning approach. *International Journal of Electrical and Computer Engineering (IJECE), 13*(4), 4504. https://doi.org/10.11591/ijece.v13i4.pp4504-4515

Huang, L., Sun, C., Qiu, X., & Huang, X. (2019). GlossBERT: BERT for word sense disambiguation with gloss knowledge. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural*

Language Processing (EMNLP-IJCNLP). https://doi.org/10.18653/v1/d19-1355

Husain, M. S. (2020). Critical concepts and techniques for information retrieval system. *Natural Language Processing in Artificial Intelligence*, 29-51. https://doi.org/10.1201/9780367808495-2

Jain, V., Chatterjee, J. M., Bansal, A., Kose, U., & Jain, A. (2022). *Computational intelligence in software modeling*. Walter de Gruyter GmbH & Co KG.

Kaddoura, S., Ahmed, R. D., & Hemanth, J. (2022). A comprehensive review on Arabic word sense disambiguation for natural language processing applications. *WIREs Data Mining and Knowledge Discovery, 12*(4). https://doi.org/10.1002/widm.1447

Kejriwal, M., Knoblock, C. A., & Szekely, P. (2021). *Knowledge graphs: Fundamentals, techniques, and applications*. MIT Press.

Krovetz, R., & Croft, W. B. (1992). Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems, 10*(2), 115-141. https://doi.org/10.1145/146802.146810

Kumar, C. S., & Santhosh, R. (2020). Effective information retrieval and feature minimization technique for Semantic Web data. *Computers & Electrical Engineering, 81*, 106518. https://doi.org/10.1016/j.compeleceng.2019.106518

Lauriola, I., Lavelli, A., & Aiolli, F. (2022). An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing, 470*, 443-456. https://doi.org/10.1016/j.neucom.2021.05.103

Lupu, M., Mayer, K., Kando, N., & Trippe, A. J. (2017). *Current challenges in patent information retrieval*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-53817-3

Matrane, Y., Benabbou, F., & Sael, N. (2023). A systematic literature review of Arabic dialect sentiment analysis. *Journal of King Saud University - Computer and Information Sciences, 35*(6), 101570. https://doi.org/10.1016/j.jksuci.2023.101570

Mesfar, S. (2010). Towards a Cascade of Morphosyntactic Tools for Arabic Natural Language Processing. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing. CICLing 2010* (Vol. 6008, pp. 150-162). Springer. https://doi.org/10.1007/978-3-642-12116-6_13

Mustafa, M., & Suleman, H. (2015). Mixed language Arabic-English information retrieval. *Computational Linguistics and Intelligent Text Processing*, 427-447. https://doi.org/10.1007/978-3-319-18117-2_32

Nagpal, M., & Petersen, J. A. (2021). Keyword selection strategies in search engine optimization: How relevant is relevance? *Journal of Retailing, 97*(4), 746-763. https://doi.org/10.1016/j.jretai.2020.12.002

Navigli, R. (2012). A quick tour of word sense disambiguation, induction and related approaches. *SOFSEM 2012: Theory and Practice of Computer Science*, 115-129. https://doi.org/10.1007/978-3-642-27660-6_10

Nguyen, Q., Vo, A., Shin, J., & Ock, C. (2018). Effect of word sense disambiguation on neural machine translation: A case study in Korean. *IEEE Access, 6*, 38512-38523. https://doi.org/10.1109/access.2018.2851281

Obeid, O., Salameh, M., Bouamor, H., & Habash, N. (2019). ADIDA: Automatic Dialect Identification for Arabic. *Paper presented at the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, Minnesota*.

Omar, A. (2021). Authorship attribution of Morsi Gameel Aziz's lyrics: A clustering-based stylometry approach. *Journal of Language and Linguistic Studies, 17*(1), 542-557. https://doi.org/10.52462/jlls.36

Omar, A., & Aldawsari, M. (2020). Lexical ambiguity in Arabic information retrieval: The case of six web-based search engines. *International Journal of English Linguistics, 10*(3), 219. https://doi.org/10.5539/ijel.v10n3p219

Omar, A., Altohami, W. M., & Hamouda, W. (2023). Exploring the efficacy and reliability of automatic text summarisation systems: Arabic texts in focus. *Cogent Arts & Humanities, 10*(1). https://doi.org/10.1080/23311983.2023.2185968

Omar, A., Elghayesh, B. I., & Kassem, M. A. M. (2019). Authorship Attribution Revisited: The Problem of Flash Fiction A morphological-based Linguistic Stylometry Approach. *Arab World English Journal, 10*(3), 318-329. https://doi:10.24093/awej/vol10no3.22

Omar, A., & Ibrahim, W. (2020). The effectiveness of stemming in the Stylometric authorship attribution in Arabic. *International Journal of Advanced Computer Science and Applications, 11*(1). https://doi.org/10.14569/ijacsa.2020.0110114

Rahman, N., & Borah, B. (2022). An unsupervised method for word sense disambiguation. *Journal of King Saud University-Computer and Information Sciences, 34*(9), 6643-6651. https://doi.org/10.1016/j.jksuci.2021.07.022

Rodd, J. (2018). Lexical ambiguity. The Oxford handbook of psycholinguistics. https://doi:10.1093/oxfordhb/9780198786825.013.5

Rubin, R. E. (2017). *Foundations of library and information science* (4th ed.). American Library Association.

Saadi, A., & Belhadef, H. (2020). Deep neural networks for Arabic information extraction. *Smart and Sustainable Built Environment, 9*(4), 467-482. https://doi.org/10.1108/sasbe-03-2019-0031

Saidi, R., Jarray, F., & Alsuhaibani, M. (2022). Comparative analysis of recurrent neural network architectures for Arabic word sense

disambiguation. *Proceedings of the 18th International Conference on Web Information Systems and Technologies*. https://doi.org/10.5220/0011527600003318

Scarlini, B., Pasini, T., & Navigli, R. (2020). With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.18653/v1/2020.emnlp-main.285

Shaalan, K., Hassanien, A. E., & Tolba, F. (2017). *Intelligent Natural Language Processing: Trends and Applications*. Springer International Publishing.

Shaalan, K., Siddiqui, S., Alkhatib, M., & Abdel Monem, A. (2018). Challenges in Arabic natural language processing. *Computational Linguistics, Speech and Image Processing for Arabic Language*, 59-83. https://doi.org/10.1142/9789813229396_0003

Sheu, P. C., Yu, H., Ramamoorthy, C. V., Joshi, A. K., & Zadeh, L. A. (2011). *Semantic computing*. John Wiley & Sons.

Simpson, A. (2019). *Language and society: An introduction*. Oxford University Press.

Senanayake, A., Gamaarachchi, H., Herath, D., & Ragel, R. (2023). DeepSelectNet: Deep neural network based selective sequencing for Oxford nanopore sequencing. *BMC Bioinformatics*, *24*(1). https://doi.org/10.1186/s12859-023-05151-0

Soudi, A., Neumann, G., & Bosch, A. V. (n.d.). Arabic computational morphology: Knowledge-based and empirical methods. *Text, Speech and Language Technology*, 3-14. https://doi.org/10.1007/978-1-4020-6046-5_1

White, R. W. (2016). *Interactions with search systems*. Cambridge University Press. https://doi.org/10.1017/CBO9781139525305

Zait, F., & Zarour, N. (2018). Addressing lexical and semantic ambiguity in natural language requirements. *2018 Fifth International Symposium on Innovation in Information and Communication Technology (ISIICT)*. https://doi.org/10.1109/isiict.2018.8613726

Zilli, A., Damiani, E., Ceravolo, P., Corallo, A., & Elia, G. (2008). *Semantic knowledge management: An ontology-based framework*. Information Science Reference.