International Journal of English Linguistics; Vol. 10, No. 3; 2020
ISSN 1923-869X E-ISSN 1923-8703
Published by Canadian Center of Science and Education

Lexical Ambiguity in Arabic Information Retrieval: The Case of Six Web-Based Search Engines

Abdulfattah Omar[1, 2] & Mohammed Aldawsari[1]

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

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

Correspondence: Abdulfattah Omar, Department of English, College of Science & Humanities, Prince Sattam Bin Abdulaziz University, Al-Kharj, Riyadh, 11942, Saudi Arabia. E-mail: a.abdelfattah@psau.edu.sa

Received: February 18, 2020 Accepted: March 25, 2020 Online Published: April 6, 2020

doi:10.5539/ijel.v10n3p219 URL: https://doi.org/10.5539/ijel.v10n3p219

Abstract

In recent years, both research and industry have shown an increasing interest in developing reliable information retrieval (IR) systems that can effectively address the growing demands of users worldwide. In spite of the relative success of IR systems in addressing the needs of users and even adapting to their environments, many problems remain unresolved. One main problem is lexical ambiguity which has negative impacts on the performance and reliability of IR systems. To date, lexical ambiguity has been one of the most frequently reported problems in the Arabic IR systems despite the development of different word sense disambiguation (WSD) techniques. This is largely attributed to the limitations of such techniques in addressing the issue of linguistic peculiarities. Hence, this study addresses these limitations by exploring the reasons for lexical ambiguity in IR applications in Arabic as one step towards reliable and practical solutions. For this purpose, the performances of six search engines *Google*, *Bing*, *Baidu*, *Yahoo*, *Yandex*, and *Ask* are evaluated. Results indicate that lexical ambiguities in Arabic IR applications are mainly due to the unique morphological and orthographic system of the Arabic language, in addition to its diglossia and the multiple colloquial dialects where sometimes mutual intelligibility is not achieved. For better disambiguation and IR performances in Arabic, this study proposes that clustering models based on supervised machine learning theory should be trained to address the morphological diversity of Arabic and its unique orthographic system. Search engines should also be adapted to the geographic location of the users in order to address the issue of vernacular dialects of Arabic. They should also be trained to automatically identify the different dialects. Finally, search engines should consider all varieties of Arabic and be able to interpret the queries regardless of the particular language adopted by the user.

Keywords: Arabic morphology, diglossia, information retrieval (IR), lexical ambiguity, word sense disambiguation (WSD)

1. Introduction

The recent years, research and industry have witnessed an increasing interest in developing reliable information retrieval (IR) systems that can effectively address the growing demands of users all over the world (Qi, Wang, & Shen, 2017; Zhang, 2016). In spite of the relative success of IR systems in addressing the needs of users and even adapting to their environments, many problems remain unresolved. One main problem is lexical ambiguity which has negative impacts on the performance and reliability of IR systems. It is even argued that lexical ambiguity is the most challenging problem for IR systems. This is because, in almost all languages, thousands of words have multiple connotations or meanings which need to be well considered in NLP applications. In English, for instance, over 80% of common English words have more than one dictionary entry, with some words having very many different definitions (Rodd, 2018). Hence, IR systems need to be trained to learn and process such words in order to achieve reliability and consistency.

As expected, recent years have witnessed the development of various techniques for resolving the problem of lexical ambiguity in (IR) applications. These have been based on what is referred to as 'word sense disambiguation' (WSD). The assumption is that determining the sense or meaning of a given word is essential for successful natural language processing (NLP) applications including IR (Agirre & Edmonds, 2007; Jacobs, 2014; Kwong, 2012; Mihalcea & Radev, 2011; Strzalkowski, 2013; Sumathy & Chidambaram, 2016; Zhekova,

2014). The WSD process is essential given that a great number of words have identical forms; moreover, they have different meanings when used in different contexts. This is technically known as polysemy. The problem with this linguistic feature, however, is that the perceived meaning of a word can vary greatly from one context to another (Ruhl, 1989). Readers/listeners, however, can quickly make use of contextual cues to select the most likely meaning when polysemous words are used within sentences and structures. Humans have the ability to reinterpret the sentence in the light of subsequent information. Evidence from brain imaging studies reveals the network of temporal and frontal brain regions that are known to be important for representing and processing ambiguous words (Rodd, 2018). It is even argued that listeners and readers rarely notice the ambiguities that pervade our everyday language (Altmann, 1998).

While it is usually easy for humans to identify the intended meaning of words with multiple meanings, it is still challenging for NLP and IR systems to determine the correct sense of such lexemes. When a word has different senses, it is difficult for the machine to determine the intended sense in a sentence (Saqib, Ahmad, Syed, Naeem, & Alotaibi, 2019; Trivedi, Sharma, & Deulkar, 2014). The word *depression* in a query, for instance, is challenging for IR systems. It is difficult for IR systems to assign its meaning to illness, weather, or economics. Thus, it is the task of WSD techniques to remove ambiguities and determine the correct sense of these words, and automatically assign the correct sense to a word with multiple meanings in a particular context (Dixit, Dutta, & Singh, 2015). The success of a given IR system depends on its ability to disambiguate, determine the correct sense, and finally retrieve only relevant documents in response to the user query.

Despite the development of different WSD techniques, evaluations of such techniques suggest that these have inherent limitations; therefore, lexical ambiguity remains the most serious problem for NLP and IR systems in Arabic. This is attributed mainly to linguistic peculiarities which are not usually considered in standard IR systems which are largely based on European languages. However, Arabic is a Semitic language, very different from European languages in terms of phonetics, morphology, syntax and semantics (Altaher, 2017; Khan & Alshara, 2019; Shaalan, Siddiqui, Alkhatib, & Abdel-Monem, 2018). Hence the challenge faced by researchers and developers of NLP applications for Arabic text and speech (Farghaly & Shaalan, 2009). It follows that IR systems should be adapted to take into consideration the unique linguistic features of Arabic.

In light of this argument, this study is undertaken in order to better understand the reasons for lexical ambiguity in the IR applications of Arabic; based on this understanding, reliable and practical solutions to the problem can then be developed. The remainder of this article is organized as follows. Section 2 surveys the main linguistic and WSD approaches for addressing the problem of lexical ambiguity in IR. Section 3 describes the methods and procedures of the study. The results of the study are reported in Section 4. Section 5 concludes this paper.

2. Literature Review

The literature suggests that the issue of lexical ambiguity has been extensively discussed in different linguistic disciplines including semantics, psycholinguistics, and discourse studies. Various semantic theories, including cognitive semantics, have been generated in order to explain the nature of lexical ambiguity and to capture as many generalizations as possible about the ambiguous and contextually-dependent nature of word meaning (Chierchia & McConnell-Ginet, 1993; Deane, 1988; Löbner, 2002; Lyons, 1975; Stallard, 1987; Tuggy, 1993). Issues of ambiguity, vagueness, polysemy, and homonymy have been the focus of lexical ambiguity studies. There is general consensus that lexical ambiguity comes from the meaning of the words, not the structure. The multiple senses of a word thus lead to more than one interpretation. Different reasons have been suggested. These include shifts in application, specialization in a social milieu, figurative language, homonyms reinterpreted, and foreign influence (Leech, 1981; Lyons, 1995). Semantic studies thus have been concerned with proposing approaches that help to determine the correct sense in ambiguous sentences. Semantic relatedness/interconnections, cognitive topology and lexical networks remain among the most popular semantic approaches to lexical ambiguity (Brugman & Lakoff, 1988).

In psycholinguistics, studies have generally focused on the mental lexicon, brain activity and responses to lexical ambiguity, and perception strategies governing the interaction between linguistic structures and performance (Durkin & Manning, 1989). Traditionally, the psycholinguistic approaches to lexical ambiguity were based one way or another on Chomsky's concept of linguistic competence. Studies in this tradition were concerned with the human ability to detect and resolve ambiguity and what an individual must know in order to comprehend and speak his language (Shultz & Pilon, 1973). In this regard, different experiments were carried out to investigate the universality of the problem. In other words, researchers sought to answer the question of whether the issue of lexical ambiguity should be considered analogous (Kess & Hoppe, 1978). This was aligned with Chomsky's concept of Universal Grammar. Under this traditional approach, lexical ambiguity was usually seen as a

disadvantage as it could result in confusion and misunderstanding. Studies in this tradition stressed that linguistic ambiguity is problematic because of its negative impact on precise language processing (Kess & Hoppe, 1981). Recent studies in psycholinguistics, however, argue that ambiguity is no longer a problem—it is something that can be taken advantage of, because easy [words] can be repeatedly used albeit in different contexts (Finn, 2012).

Interestingly, in both semantics and psycholinguistics, discourse-based approaches have been used in the investigation of lexical ambiguity. In semantics, discourse is suggested as a mechanism for the resolution of lexical ambiguity. The focus is no longer on semantic relatedness. Likewise, the integration of discourse was tested and proved effective in helping individuals with aphasia and brain damage to resolve lexical ambiguity (Mason & Just, 2007; Tompkins, Baumgaertner, Lehman, & Fassbinder, 2000).

With the development of computational theory and NLP studies, the issue of lexical ambiguity has once again been the focus of many researchers. Different techniques have been developed in recent years to address the problem of lexical ambiguity and improve the performance of IR systems. Work on lexical ambiguity has traditionally focused on developing WSD techniques. The assumption has been that there is a close relationship between WSD and the IR. Therefore, correct disambiguation of words can lead to improvements in the effectiveness of retrieval systems (Sanderson, 1994; Zhong & Ng, 2012). Determining the correct sense or meaning of a given word increases the potential of IR systems to suggest relevant documents for a given user query.

According to the literature, there are three main WSD approaches: dictionary-based, ontology-based, and knowledge-based. The dictionary-based approach is usually considered to be the traditional WSD method and it is based on the development of corpus-based studies that use electronic corpora to resolve ambiguity issues. In this approach, a word's meanings are compared to those of the surrounding text where all the senses of a word that need to be disambiguated are retrieved from the dictionary (Agirre & Edmonds, 2007; Chen, 2000; Pal & Saha, 2015; Zhekova, 2014). One of the earlier attempts to implement this approach was Lesk's (1986) use of *Oxford's Advanced Learner's Dictionary of Current English* to resolve the issue of word senses (Indurkhya & Damerau, 2010). Similarly, Guthrie et al. used the *Longman Dictionary of Contemporary English* in 1991 to remove ambiguities and identify the correct sense of polysemous entries through the use of subject codes (Pal & Saha, 2015). The underlying principle in this approach is that there is a set of complete entries for each polysemous expression, from which anomalous alternatives are subsequently eliminated and only relevant senses are retained. Despite the continued research on dictionary-based approaches and techniques, lexical ambiguity remains pervasive so that many doubts have been raised about the reliability of these methods (Agirre & Edmonds, 2007). One major problem with this approach is that it is based on what can be described as 'static knowledge' as it makes no use of any specific knowledge manipulation mechanisms apart from the simple ability to match valences of structurally-related words (Boguraev & Pustejovsky, 1990).

With knowledge-based techniques, the main assumption is that disambiguation systems need sources of knowledge to determine the proper meaning of a lexeme that has multiple senses (Otegi, Arregi, Ansa, & Agirre, 2015; Sheng, Fan, Thomas, & Ng, 2001). Hence, these approaches are similar to dictionary-based ones in that both rely on sources of knowledge for disambiguation purposes. However, dictionary-based techniques are limited to the use of dictionaries, whereas knowledge-based techniques exploit different sources such as specialized corpora, WordNet and semantic systems. It is through these sources of knowledge that WSD systems are able to disambiguate words by means of defining their contexts. In other words, corpora, WordNet, and other sources of knowledge are used as the contexts for disambiguating lexemes with multiple senses. One major problem with knowledge-based approaches, however, is that they are based only on words to disambiguate target words. Devendra and Salakhutdinov (Chaplot & Salakhutdinov, 2018) explain that the sense of a word depends on not just the words in the context but also on their senses. Since the senses of the words in the context are also unknown, they need to be optimized jointly.

In order to overcome the limitations of both dictionary-based and knowledge-based approaches, ontology-based techniques have been developed. Ontologies are the most widely-used techniques in IR systems. In the ontology-based approach, words with multiple senses are disambiguated through the design of ontology of semantic concepts. The function of this ontology is to enable IR systems to resolve lexical ambiguity problems by drawing inferences from the concept network of the ontology (Hadzic, Chang, & Wongthongtham, 2009; Ławrynowicz, 2017; Mena & Illarramendi, 2001). The underlying principle of ontology-based techniques has been that searches in IT should be based on meaning and inference rather than on literal strings. IR systems and search engines should be equipped with mechanisms enabling them to understand the relationship between search items and concepts. However, in spite of their advantages in terms of enriching semantic inference and expressiveness, making inferences and understanding relationships between search items, deep levels of

conceptualization are necessary. These usually require complex query languages that are not really appropriate in many cases (Sy et al., 2012).

3. Methodology

In order to understand the reasons for lexical ambiguity in the Arabic IR systems, in this study, the performances of six search engines are evaluated. These are *Google*, *Bing*, *Baidu*, *Yahoo*, *Yandex*, and *Ask*. These are ranked as the most popular search engines in the world according to the Digital Marketing Agency (Chris, 2019). Various criteria are applied when evaluating IR systems. These include the ability to retrieve relevant documents, the ability to avoid unwanted and irrelevant documents, the response time, the way retrieval performance is fulfilled and achieved, and the techniques for improving performance effectiveness (Büttcher, Clarke, & Cormack, 2016; Goker & Davies, 2009; Harman & Marchionini, 2011; Lupu, Mayer, Kando, & Trippe, 2017; Strzalkowski, 2013; White, 2016). The evaluation of the performance of IR systems is depicted in Figure 1.

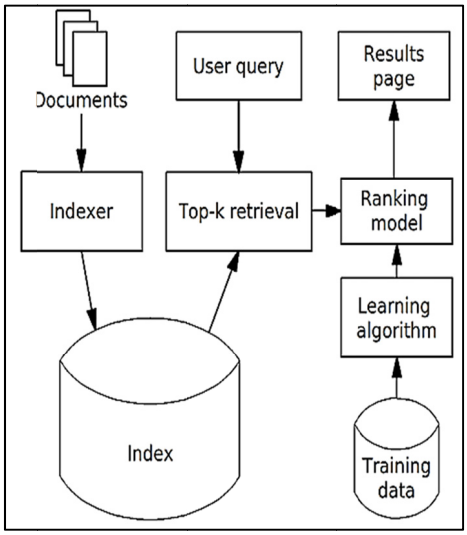

Figure 1. Evaluation of IR systems

This study, however, is limited to the identification of the ways that IR systems address the issue of lexical ambiguity. Here, we are concerned with investigating the ability of IR systems to retrieve only wanted items. In other words, we are mainly concerned with the precision of retrieval. This is the ability of an IR system to retrieve only those documents relevant to a query. Therefore, IR performance is measured by calculating the ratio of relevant documents being retrieved by the system to the total number of retrieved documents as shown in the following example:

Total number of retrieved documents = 100

Number of only relevant documents = 45

$$\text{Precision} = \frac{45}{100} \text{x } 100$$

This means that Precision = 45%.

Given a query ‘What is lexical ambiguity?’, we use Google to show how precision is measured. As seen in Figure 2, the total number of retrieved documents is 2,810,000 items. Assuming that only 1,405,000 are relevant documents, this means that precision is 50%.

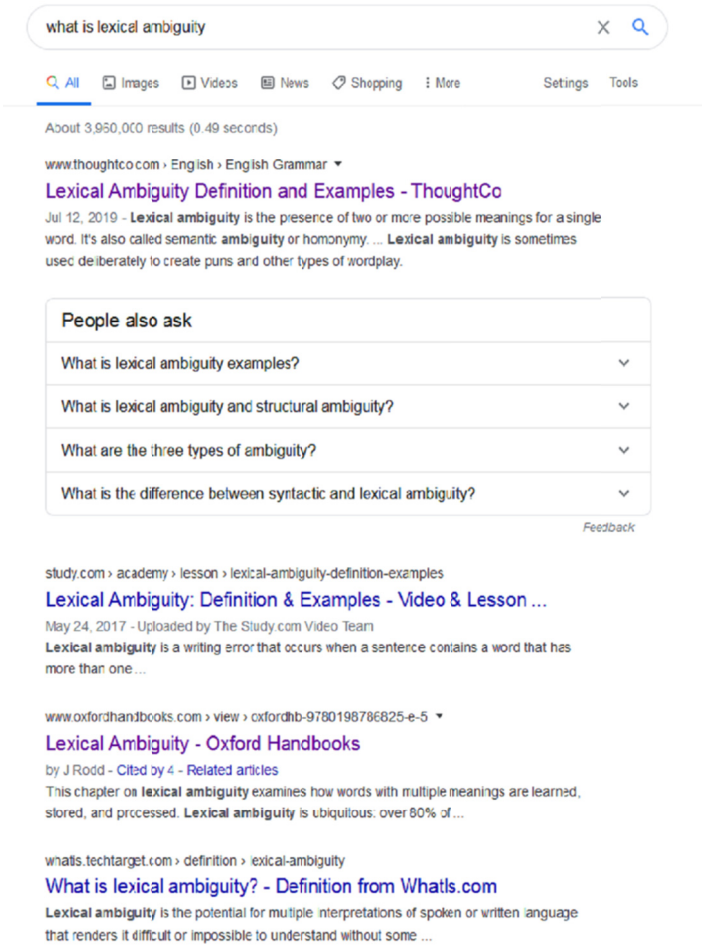

Figure 2. An illustration of Google's retrieval performance

In order to evaluate the recall performance of the six selected search engines and identify the sources of lexical ambiguity, five queries were used. These were randomly selected from the most popular topics on search engines and Google trends in Arab countries in 2019. These are shown below in Table 1.

Table 1. List of the queries

Query	English translation
الكان 2019	Africa Cup of Nations 2019
الصوفية	Sufism
سبق	Sapq
قرآن	Quran (the Holy Book of Muslims)
مرسي علم	Marsa Alam (Coastal city in Egypt)

The next step is to investigate the results for the above queries, identify irrelevant results due to lexical ambiguities, and finally suggest the reasons for lexical ambiguity in the performance of the selected search engines.

4. Analysis and Discussions

Results indicate that there are significant differences in terms of the retrieval performance. For all the queries, Google is ranked the first. However, this study is not primarily concerned with ranking the performance of the selected search engines. It is mainly concerned with investigating the performance of the selected search engines by exploring reasons of lexical ambiguity in Arabic IR.

The investigation led to the conclusion that lexical ambiguity is the main reason that irrelevant items were selected in response to the selected queries. Overall, lexical ambiguity can be grouped under three main categories: the unique morphological and orthographic system, the diglossia feature, and the multiple colloquial dialects. These represent real challenges for IR systems and have negative impacts on their performance as explained below.

Results indicate that thousands of irrelevant documents were generated due to the unique morphological features which are not taken into account by the search engines. The Arabic language has a unique morphological system which can lead to an incorrect meaning being assigned to a particular word. This can be explained as follows. In order to determine the sense or meaning of a word, the three-letter root must be identified, followed by the identification of the syntactic context (Akesson, 2010; Ryding, 2005; Soudi, van den Bosch, & Neumann, 2007). However, in some cases, its meaning can still be ambiguous, and will need to be disambiguated (Glanville, 2018; Habash, 2010; Ryding, 2014). That is, it is sometimes difficult to relate the meaning of a given word to its three-letter root. The word مسكين(poor) as in يا له من ولد مسكين(What a poor guy), for instance, has no connection to the three-letter form سكن (literally translated as being constant or inhabited). This is partly due to the inevitable evolution of Arabic, just as in any other language. Hence, very often it is difficult for those IR systems based on Arabic dictionaries and glossaries to determine the sense or meaning of a given word. Additionally, Arabic is a synthetic language that is based on the case system. This case system is not usually used by Arabic people in spite of its importance in determining the correct meaning or sense of the word as shown in Table 2.

Table 2. An illustration of the case system in Arabic

Arabic form	Part of speech	English translation
عِقْد	Noun	Necklace
عِقْد	Noun	Decade
عَقْد	Noun	Contract
عَقَدَ	Verb	Held
عَقَّدَ	Verb	Complicated
عُقَد	Noun	Knots

Generally, Internet users are not familiar with the use of cases in their search. Furthermore, the vast majority of Arabic texts are not written using the case system. This poses real challenges for search engines and IR systems that are attempting to retrieve only relevant documents or items in response to users' queries in Arabic.

Another reason for the lexical ambiguity in Arabic is the feature of diglossia, of which there are two types. These are Modern Standard Arabic (MSA) which is considered the H (High) variety and Colloquial Arabic which are classified as the L (Low) variety. In the Arab countries, MSA is the official language and the formal language of education in schools. It is also used in the Press and TV news bulletins. Educated Arab speakers are usually able to produce and understand MSA, while uneducated people usually have difficulties in producing and even understanding this variety of Arabic (Albirini, 2016; Ferguson, 1996; Owens, 2013). There are great similarities between MSA and Classical Arabic (the language of the Quran and classical literature) especially in terms of morphology, grammar and structure. However, although MSA follows the basic syntax and morphology of Classical Arabic, the vocabulary is widely different (Ibrahim, 2009; Simpson, 2019). Colloquial Arabic, in turn, refers to the regional vernacular dialects. It is the language used in everyday speech (AlSuwaiyan, 2018). It is an umbrella term that covers various Arabic dialects including Egyptian Colloquial Arabic, Lebanese Colloquial Arabic, and Moroccan Colloquial Arabic. The morphological, lexical, and grammatical features of CA are very different from those of MSA (Bassiouney, 2009). Many words in MSA are used differently in CA, making it difficult for IR systems and search engines to determine the correct sense. It was also observed that the significant changes in the vernacular dialects of CA represent a real challenge to the performance of IR systems. Although these vernacular dialects of Arabic were not written and for centuries had been used only for oral communication, they are now widely used in writing, especially with the development of communication technologies, the proliferation of social media platforms, and the increasing interaction between people (Bassiouney, 2009; Harrat, Meftouh, & Smaili, 2019; Khedher et al., 2015).

The results of this study align with those reported in the literature in that the reasons for lexical ambiguity are not the same for all natural languages. This suggests that the linguistic peculiarities of a particular language should be considered by IR engineers if they are to provide workable and reliable solutions for the problem of lexical ambiguity (Dini L. & V., 1999; Kraaij, 2004; Mustafa & Suleman, 2015). Furthermore, all variations of Arabic

must be taken into account during the development of IR systems. The colloquial Arabic dialects have long been ignored in NLP and IR applications, with the current search engines still catering mostly to MSA (Azmi & Aljafari, 2015; Obeid, Salameh, Bouamor, & Habash, 2019). IR systems are generally trained to deal with Standard Arabic which is in many ways different from the Arabic colloquial dialects. Thus, it is imperative that IR systems and search engines integrate these colloquial dialects to address the day-to-day needs of users all over the world. CA is the primary language of communication and younger generations are more adept at communicating in CA (Azmi & Aljafari, 2015; Bassiouney, 2009).

For better disambiguation and IR system performance in terms of Arabic, this study proposes that clustering models based on supervised machine learning theory should be trained to address the morphological diversity of the Arabic language and its unique orthographic system. Search engines should also be adapted to the geographic location of the users in order to address the issue of Arabic vernacular dialects. They should also be trained to automatically identify the various dialects, which will lead to the improvement in the IR performance as it reduces the possibility of having words with multiple meanings (Obeid et al., 2019; Sadat, Kazemi, & Farzindar, 2014).

5. Conclusion

In this article, we explored the reasons for lexical ambiguity in Arabic IR systems in order as a first step to proposing reliable and workable WSD solutions. It was revealed that linguistic peculiarities have important implications for IR engineering and performance. In Arabic, these have an impact on the reliability of IR systems and search engines. There are serious limitations of the selected search engines in considering the linguistic peculiarities of Arabic which constitute the main reasons for linguistic ambiguity in Arabic IR. These can be mainly attributed to the unique morphological system of Arabic, its diglossia, and the numerous colloquial dialects. WSD techniques need to consider these linguistic peculiarities for a better IR system performance. This paper was limited to considering the use of only the Arabic alphabet by search engines. Future work can focus on lexical ambiguity in the emerging Arabic chat Alphabets usually referred to as Franco-Arabic or Arabizi.

6. Acknowledgments

We take this opportunity to thank Prince Sattam Bin Abdulaziz University in Saudi Arabia alongside its Deanship of Scientific Research, for all the technical support it has unstintingly provided for the fulfillment of the current research project.

References

Agirre, E., & Edmonds, P. (2007). *Word Sense Disambiguation: Algorithms and Applications*. Springer Netherlands. https://doi.org/10.1007/1-4020-4809-2

Akesson, J. (2010). *The Basics & Intricacies of Arabic Morphology*. Pallas Athena Distribution.

Albirini, A. (2016). *Modern Arabic Sociolinguistics: Diglossia, Variation, Codeswitching, Attitudes and Identity*. Taylor & Francis. https://doi.org/10.4324/9781315683737

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

Altaher, A. (2017). Hybrid approach for sentiment analysis of Arabic tweets based on deep learning model and features weighting. *International Journal of Advanced and Applied Sciences*, *4*(8), 43–49. https://doi.org/10.21833/ijaas.2017.08.007

Altmann, G. T. M. (1998). Ambiguity in sentence processing. *Trends in Cognitive Science*, *2*(4), 144–152. https://doi.org/10.1016/S1364-6613(98)01153-X

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

Bassiouney, R. (2009). *Arabic Sociolinguistics: Topics in Diglossia, Gender, Identity, and Politics*. Georgetown University Press. https://doi.org/10.3366/edinburgh/9780748623730.001.0001

Boguraev, B., & Pustejovsky, J. (1990). *Lexical Ambiguity and The Role of Knowledge Representation in Lexicon Design*. Paper presented at the 13th International Conference on Computational Linguistics, COLING. https://doi.org/10.3115/997939.997946

Brugman, C., & Lakoff, G. (1988). Cognitive Topology and Lexical Networks1. In S. Small, G. Cottrell & M. Tanenhaus (Eds.), *Lexical Ambiguity Resolution: Perspective from Psycholinguistics, Neuropsychology and Artificial Intelligence* (pp. 477–508). https://doi.org/10.1016/B978-0-08-051013-2.50022-7

Büttcher, S., Clarke, C. L. A., & Cormack, G. V. (2016). *Information Retrieval: Implementing and Evaluating Search Engines*. MIT Press.

Chaplot, D. S., & Salakhutdinov, R. (2018). *Knowledge-based Word Sense Disambiguation using Topic Models*. AAAI.

Chen, J. N. (2000). Adaptive Word Sense Disambiguation Using Lexical Knowledge in a Machine-readable Dictionary. *Computational Linguistics and Chinese Language Processing*, *5*(2), 1–42.

Chierchia, G., & McConnell-Ginet, S. (1993). *Meaning and Grammar: An introduction to Semantics*. Cambridge: Messachusetts MIT Press.

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

Deane, P. (1988). Polysemy and Cognition. *Lingua*, *75*. https://doi.org/10.1016/0024-3841(88)90009-5

Dini L., D., & V., T. (1999). Linking Theory and Lexical Ambiguity: The Case of Italian Motion Verbs. In B. Harry & M. Reinhard (Eds.), *Computing Meaning* (Part of the Studies in Linguistics and Philosophy book series, Vol. 73, pp. 321–337). Dordrecht: Springer. https://doi.org/10.1007/978-94-011-4231-1_16

Dixit, V., Dutta, K., & Singh, P. (2015). Word Sense Disambiguation and Its Approaches. *CPUH-Research Journal*, *1*(2), 54–58.

Durkin, K., & Manning, J. (1989). Polysemy and the subjective lexicon: Semantic relatedness and the salience of intra-word senses. *Journal of Psycholinguistic Research*, *18*, 577–612. https://doi.org/10.1007/BF01067161

Farghaly, A., & Shaalan, K. F. (2009). Arabic Natural Language Processing: Challenges and Solutions. *ACM Transactions on Asian Language Information Processing*, *8*(4). https://doi.org/10.1145/1644879.1644881

Ferguson, C. (1996). Epilogue: Diglossia Revisited. In E.-S. M. Badawi & A. Elgibali (Eds.), *Understanding Arabic: Essays in Contemporary Arabic Linguistics* (pp. 49–67). Cairo: American University Press.

Finn, E. (2012, January 19). *The advantage of ambiguity*. MIT News.

Glanville, P. J. (2018). *The Lexical Semantics of the Arabic Verb*. Oxford University Press. https://doi.org/10.1093/oso/9780198792734.001.0001

Goker, A., & Davies, J. (2009). *Information Retrieval: Searching in the 21st Century*. Wiley.

Habash, N. Y. (2010). *Arabic Natural Language Processing*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00277ED1V01Y201008HLT010

Hadzic, M., Chang, E. J., & Wongthongtham, P. (2009). *Ontology-Based Multi-Agent Systems*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-01904-3

Harman, D. K., & Marchionini, G. (2011). *Information Retrieval Evaluation*. Morgan & Claypool.

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

Ibrahim, Z. (2009). *Beyond Lexical Variation in Modern Standard Arabic: Egypt, Lebanon and Morocco*. Cambridge Scholars Publisher.

Indurkhya, N., & Damerau, F. J. (2010). *Handbook of Natural Language Processing*. CRC Press. https://doi.org/10.1201/9781420085938

Jacobs, P. S. (2014). *Text-based intelligent Systems: Current Research and Practice in information Extraction and Retrieval*. Taylor & Francis. https://doi.org/10.4324/9781315806952

Kess, J. F., & Hoppe, R. A. (1978). On psycholinguistic experiments in ambiguity. *Lingua*, *45*(2), 125–140. https://doi.org/10.1016/0024-3841(78)90002-5

Kess, J. F., & Hoppe, R. A. (1981). *Ambiguity in Psycholinguistics*. London: John Benjamins Publishing Company. https://doi.org/10.1075/pb.ii.4

Khan, S., & Alshara, M. (2019). Development of Arabic evaluations in information retrieval. *International Journal of Advanced and Applied Sciences*, *6*(12), 92–98. https://doi.org/10.21833/ijaas.2019.12.011

Khedher, M. Z., Abandah, G. A., Al-Anati, W. A., Ababneh, S. M., Zghoul, A. A., & Hattab, M. S. (2015). *Effect of topic on the Arabic language used on social networks and mobile phone communications*. Paper presented at the 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing

Technologies (AEECT), 3–5 November. https://doi.org/10.1109/AEECT.2015.7360593

Kraaij, W. (2004). *Variations on Language Modeling for Information Retrieval*. University of Twente, AE Enschede, The Netherlands.

Kwong, O. Y. (2012). *New Perspectives on Computational and Cognitive Strategies for Word Sense Disambiguation*. Springer New York. https://doi.org/10.1007/978-1-4614-1320-2

Ławrynowicz, A. (2017). *Semantic Data Mining: An Ontology-based Approach*. IOS PRESS.

Leech, G. (1981). *Semantics: The Study of Meaning*. Middlesex: Penguin Books.

Löbner, S. (2002). *Understanding Semantics*. London Arnold.

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

Lyons, J. (1975). *Introduction to Theoretical Linguistics*. Cambridge: Cambridge University Press.

Lyons, J. (1995). *Linguistic Semantics: An Introduction*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511810213

Mason, R. A., & Just, M. A. (2007). Lexical ambiguity in sentence comprehension. *Brain Research*, *1146*, 115–127. https://doi.org/10.1016/j.brainres.2007.02.076

Mena, E., & Illarramendi, A. (2001). *Ontology-Based Query Processing for Global Information Systems*. Springer US. https://doi.org/10.1007/978-1-4615-1441-1

Mihalcea, R., & Radev, D. (2011). *Graph-based Natural Language Processing and Information Retrieval*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511976247

Mustafa, M., & Suleman, H. (2015). Mixed Language Arabic-English Information Retrieval. In G. A. (Ed.), *Computational Linguistics and Intelligent Text Processing*. CICLing 2015. Cham Springer. https://doi.org/10.1007/978-3-319-18117-2_32

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.

Otegi, A., Arregi, X., Ansa, O., & Agirre, E. (2015). Using knowledge-based relatedness for information retrieval. *Knowledge and Information Systems*, *44*(3), 689–718. https://doi.org/10.1007/s10115-014-0785-4

Owens, J. (2013). *The Oxford Handbook of Arabic Linguistics*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199764136.001.0001

Pal, A. R., & Saha, D. (2015). Word Sense Disambiguity: A Survey. *International Journal of Control Theory and Computer Modeling*, *5*(3), 1–16. https://doi.org/10.5121/ijctcm.2015.5301

Qi, A., Wang, Y., & Shen, C. (2017). Application of Courseware Based on Information Retrieval Technology. *International Journal of Emerging Technologies in Learning*, *11*(3), 32–36. https://doi.org/10.3991/ijet.v11i03.5346

Rodd, J. (2018). Lexical Ambiguity. In S.-A. Rueschemeyer & M. G. Gaskell (Eds.), *The Oxford Handbook of Psycholinguistics*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198786825.013.5

Ruhl, C. (1989). *On monosemy: A study in linguistic semantics*. Albany: State University of New York.

Ryding, K. C. (2005). *A Reference Grammar of Modern Standard Arabic*. Cambridge University Press. https://doi.org/10.1017/CBO9780511486975

Ryding, K. C. (2014). *Arabic: A Linguistic Introduction*. Cambridge University Press. https://doi.org/10.1017/CBO9781139151016

Sadat, F., Kazemi, F., & Farzindar, A. (2014). *Automatic Identification of Arabic Language Varieties and Dialects in Social Media* (pp. 22–27). Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP). https://doi.org/10.3115/v1/W14-5904

Sanderson, M. (1994). Word Sense Disambiguation and Information Retrieval. In B. W. Croft & C. J. van Rijsbergen (Eds.), *SIGIR '94*. London: Springer. https://doi.org/10.1007/978-1-4471-2099-5_15

Saqib, S. M., Ahmad, S., Syed, A. H., Naeem, T., & Alotaibi, F. M. (2019). Analysis of latent Dirichlet allocation

and non-negative matrix factorization using latent semantic indexing. *International Journal of Advanced and Applied Sciences*, *6*(10), 94–102. https://doi.org/10.21833/ijaas.2019.10.015

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

Sheng, F., Fan, X., Thomas, G., & Ng, P. (2001). A Knowledge-Based Approach to Effective Document Retrieval. *Journal of Systems Integration*, *10*, 411–436. https://doi.org/10.1023/A:1011262119636

Shultz, T. R., & Pilon, R. (1973). Development of the Ability to Detect Linguistic Ambiguity. *Child Development*, *44*(4), 728–733. https://doi.org/10.2307/1127716

Simpson, A. (2019). *Language and Society: An Introduction*. Oxford University Press.

Soudi, A., van den Bosch, A., & Neumann, G. (2007). *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer Netherlands. https://doi.org/10.1007/978-1-4020-6046-5

Stallard, D. (1987). *The Logical Analysis of Lexical Ambiguity*. ACL. https://doi.org/10.3115/981175.981200

Strzalkowski, T. (2013). *Natural Language Information Retrieval*. Springer Netherlands.

Sumathy, K. L., & Chidambaram, M. (2016). An Advanced ontology based automatic approach in improving the similarity by means of combining the sub‐graphs between the information. *International Journal of Advanced and Applied Sciences*, *3*(8), 1–6. https://doi.org/10.21833/ijaas.2016.08.001

Sy, M.-F., Ranwez, S., Montmain, J., Regnault, A., Crampes, M., & Ranwez, V. (2012). User centered and ontology-based information retrieval system for life sciences. *BMC Bioinformatics*, *13*, 1–12. https://doi.org/10.1186/1471-2105-13-S1-S4

Tompkins, C. A., Baumgaertner, A., Lehman, M. T., & Fassbinder, W. (2000). Mechanisms of discourse comprehension impairment after right hemisphere brain damage: Suppression in lexical ambiguity resolution. *Journal of Speech, Language, and Hearing Research*, *43*(1), 62–78. https://doi.org/10.1044/jslhr.4301.62

Trivedi, M., Sharma, S., & Deulkar, K. (2014). Approaches to Word Sense Disambiguation. *International Journal of Engineering Research & Technology*, *3*(10), 645–647.

Tuggy, D. (1993). Ambiguity, polysemy, and vagueness. *Cognitive Linguistics*, *4*(3), 273–290. https://doi.org/10.1515/cogl.1993.4.3.273

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

Zhang, L. (2016). Study on the Application of Web Information Retrieval in the Teaching of Language Translation. *International Journal of Emerging Technologies in Learning*, *11*(4), 114–119. https://doi.org/10.3991/ijet.v11i04.5550

Zhekova, D. (2014). *Automatic Extraction of Examples for Word Sense Disambiguation*. GRIN Verlag.

Zhong, Z., & Ng, H. T. (2012). *Word Sense Disambiguation Improves Information Retrieval* (vol. 1, pp. 273–282). Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics.