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

Abdulfattah Omar¹,² & Mohammed Aldawsari¹

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,
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 language deficit. Traditional psycholinguistic approaches to lexical ambiguity were based on the assumption that lexical ambiguity comes from the structure of the language itself (Altmann, 1998). Traditional approaches 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 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; Ławryniewicz, 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](image)

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:

\[
\text{Precision} = \frac{\text{Number of only relevant documents}}{\text{Total number of retrieved documents}} \times 100
\]

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%.
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

<table>
<thead>
<tr>
<th>Arabic Query</th>
<th>English Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>العالم (Coastal City)</td>
<td>Marsa Alam (Coastal city in Egypt)</td>
</tr>
<tr>
<td>سفريه (the Holy Book)</td>
<td>Quran (the Holy Book of Muslims)</td>
</tr>
<tr>
<td>LOOKS (the Holy Book)</td>
<td>Quran (the Holy Book of Muslims)</td>
</tr>
<tr>
<td>COASTAL CITY</td>
<td>Marsa Alam (Coastal city in Egypt)</td>
</tr>
<tr>
<td>Africa Cup of Nations</td>
<td>Africa Cup of Nations 2019</td>
</tr>
<tr>
<td>Sufism</td>
<td>Sufism</td>
</tr>
<tr>
<td>Sapq</td>
<td>Sapq</td>
</tr>
</tbody>
</table>

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 
miskin
(poor) as in
laya lawd miskin (What a poor guy), for instance, has no connection to the three-letter form
sken (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

<table>
<thead>
<tr>
<th>Arabic form</th>
<th>Part of speech</th>
<th>English translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>عِقْد</td>
<td>Noun</td>
<td>Necklace</td>
</tr>
<tr>
<td>عِقْد</td>
<td>Noun</td>
<td>Decade</td>
</tr>
<tr>
<td>عَقْد</td>
<td>Noun</td>
<td>Contract</td>
</tr>
<tr>
<td>عَقَدَ</td>
<td>Verb</td>
<td>Held</td>
</tr>
<tr>
<td>عُقَد</td>
<td>Verb</td>
<td>Complicated</td>
</tr>
<tr>
<td>عَقﱠدَ</td>
<td>Noun</td>
<td>Knots</td>
</tr>
</tbody>
</table>

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, Mefouh, & 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


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


Copyrights

Copyright for this article is retained by the author, with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).