Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6, 4700-4705
2024
Publisher: Learning Gate
DOI: 10.55214/25768484.v8i6.3018

Challenges in natural Arabic language processing

Yazeed Al Moaiad[1*], Mohammad Alobed[2], Mahmoud Alsakhnini[3], Alaa M. Momani[4]
[1,2]Faculty of Computer and Information, Al-Madinah International University, Kuala Lampur, Malaysia; yazeed.alsayed@mediu.edu.my (Y.A.M.) shamona.2018@gmail.com (M.A.).
[3,4]School of Computing, Skyline University College, Sharjah, United Arab Emirates; m.alsakhnini@gmail.com (M.A.) momanialaa@gmail.com (A.M.).

Abstract: Speech recognition and text summarization, plagiarism detection, machine translation, chatbots, sentiment analysis (SA), question answering (QA), and dialogue systems are all products of natural language processing (NLP), a branch of AI concerned with modeling natural languages for the purpose of developing relevant applications. NLP draws on several disciplines, not only computer science and linguistics, for its research and development. These include cognitive science, psychology, mathematics, and more. More than 1.5 billion Muslims throughout the world depend on the Arabic language for their daily five-times-prayer practice, and Arabic is one of six official languages used by more than 422 million people in the Arab world, according to UNESCO. Dialectal Arabic is the slang language spoken informally in everyday life and varies from country to country; Classical Arabic is a reflection of the language spoken by the Arabs more than fourteen centuries ago. Modern Standard Arabic is an evolving variety of Arabic that borrows and innovates regularly to meet the changing needs of its speakers. Complexity is added to the Arabic language by the fact that it encompasses not one but three different varieties of spoken language: classical, contemporary, and colloquial. However, Arabic language processing on computers remains difficult for a variety of reasons. Because of the extensive inflection and derivational processes that occur in Arabic, a single lemma may be utilized to produce several words with distinct meanings.

Keywords: *Arabic dialects, Morphological richness, Orthographic Ambiguity, Orthographic inconsistency, Resource poverty, Stemming.*

1. Introduction

There are several obstacles to overcome when attempting to apply NLP techniques to the Arabic language. Arabic has a complex morphology and is also notoriously vague. Words in Modern Standard Arabic MSA typically undergo 12 morphological analyses, which is much more than the average of 4 in English, and a full set of part-of-speech tags in MSA has over 300,000 tags (English has approximately 50). (English has 1.25 POS tags per word on average). Because Arabic orthography nearly never uses diacritics to indicate short vowels and double consonants, there is a significant degree of ambiguity. Many Arabic dialects exist, and they all vary significantly from Modern Standard Arabic (MSA), in the same way as romance languages diverge from Latin. Although it is the standard version of Arabic, Modern Standard Arabic (MSA) is not a spoken language among the Arab population. The actual mother languages are spoken dialects that have no written norms and little linguistic resources [1].

Many initiatives have been made over the last two decades to build tools to aid with Arabic natural language processing (NLP). While some of these tools are designed to do one or a few particular NLP tasks, others make an effort to accomplish many jobs simultaneously.

ABU DHABI, UAE— (BUSINESS WIRE)— Technology Innovation Institute (TII), a global research center and applied research arm of Abu Dhabi's Advanced Technology Research Council, has announced the launch of NOOR, the world's largest Arabic natural language processing (NLP) model to date. As digital technologies like as chatbots, market intelligence, and machine translation lean strongly toward English and Chinese-speaking markets, the Technology Innovation Institute's Noor model may

* Correspondence: yazeed.alsayed@mediu.edu.my

provide the Arab world a fresh advantage in the quest to digitalize. NOOR's training dataset combines online data with novels, poetry, news stories, and technical knowledge to considerably broaden the model's application, making it the biggest high-quality cross-domain Arabic dataset in existence. Model has 10 billion parameters, most in Arabic [2].
Arabic and Arabic dialect pre-processing, morphological modeling, dialect identification, named object recognition, and sentiment analysis are all supported by CAMeL Tools, an open-source Python toolkit. The CAMeL Tools package includes APIs and command-line interfaces (CLIs) for accessing these tools programmatically [3].

Pre-processing is the process of cleaning and preparing data for analysis. Document strings are broken into smaller chunks during the tokenization pre-processing step. A stop word is a term that has no significance or importance in the context of the search. After the text has been transformed to a token list, stop words are eliminated [3]. Stemming is a way of creating verb stems or roots by either utilizing a dictionary to extract affixes associated to its root or by establishing a set of linguistic rules to recognize verb patterns and thus extracting the root. There are two basic stemming techniques in Arabic: the root-based approach and the light stemming approach. Light stemming methods merely extract prefixes and suffixes from phrases, while root stemming algorithms remove all prefixes, suffixes, and infixes [4].

Arabic, as comparison to English, has less tools available for natural language processing and related subjects, and is one of the low resources languages, according to the Association for Computational Linguistics (ACM). Researchers in low-resource languages, such as Arabic, have less resources at their disposal than their English-speaking counterparts [5].

2. Literature Review

2.1. Arabic NLP Applications

The following are the several NLP initiatives that address these difficulties for Arabic, organized by subcategory:

2.1.1. An Automatic Arabic Text Summarization

Text summary creates a compacted version of a document with key information. Material summary reduces the original text into a shorter version that retains key points. Extractive and abstractive summarization are two methods. Extractive summarization selects the most important sentences. Abstract summarization creates unique sentences from the source material. The authors describe an extractive-based single-document strategy for Arabic text summarization using Genetic Algorithms [5].

2.1.2. Plagiarism Detection

Plagiarism is by no means a new problem in the academic world; however, with the proliferation of the internet and the resulting ease with which one can gain access to content from around the world, the problem has taken on a greater significance, rendering human intervention alone insufficient to combat it. Despite this, plagiarism is far from being an unsolved issue, since computer-assisted plagiarism detection is now an active topic of study that comes under the umbrella of the fields of Information Retrieval (IR) and Natural Language Processing (NLP) [6].

2.1.3. Arabic Chatbot

A chatbot may use natural language to speak in real time. Due to the Internet development, more companies are employing chatbots to enhance customer service, simplify procedures, and boost productivity. Chatbots are in use, but organizations want autonomous, conversational agents. In entertainment, medical, education, and business, chatbots have proved effective in English, Spanish, and French. Arabic chatbots, particularly in maintenance, are difficult and rare. Chatbots have shown to be effective in a variety of industries, including entertainment, medical, education, and business, for various languages, including English, Spanish, and French. However, Arabic chatbots are difficult to find and uncommon, particularly in the support arena [7].

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 4700-4705, 2024
DOI: 10.55214/25768484.v8i6.3018

2.1.4. An Automated Arabic Essay Scoring

One of the most significant uses of natural language processing is something called automated essay scoring, sometimes known as AES (NLP). It would appear that there is an immediate need for automated scoring systems in order to evaluate the students' responses to Arabic essay questions. This is primarily attributable to the rapid growth of the educational community as well as the increased amount of time and effort required to complete traditional examinations or assessment scoring procedures. Feedback of score is produced by AES based on the text similarity approaches in order to decrease the amount of time and effort required. Text-to-text similarity is determined by comparing the texts of the student's response and the model answer provided by the teacher using a variety of comparative methodologies to assess the level of similarity between the two [8, 9, 10].

2.1.5. Arabic Machine Translation

In recent years, the computer language field has seen significant evolution, with applications in a variety of disciplines. Machine Translation (MT) technology has advanced significantly as a discipline, with many methods and methodologies. Although there are several MT systems and tools that support Arabic, the translation quality is modest and should be improved. Furthermore, the increased need for effective technologies to process and translate information from/to Arabic encouraged Arabic Machine Translation (AMT) researchers to suggest new techniques and solutions based on the mainstream method, particularly neural machine translation (NMT) [11].

2.2. Challenges of Processing the Arabic Language

To model modern Arabic is difficult for NLP for a variety of reasons, including the language's morphological depth, orthographic ambiguity, dialectal variances, orthographic noise, and lack of available resources. Right-to-left Arabic typography is a non-issue since it has been satisfactorily resolved (although not universally implemented).

2.2.1. Various Forms of Letters

Depending on the context in which they appear in a word, Arabic letters may be written in a wide variety of different ways. For instance, the letter "ف/faa" may appear like "فـ/ فـ / ف" depending on where it is in the word, whether it is at the beginning, in the middle, or at the conclusion of the sentence [12, 13]. Take, for example:

- Cat/ هرة - begging of the word
- Plain/ سهل -Middle of the word
- Water/مياه -End of the word

2.2.2. Orthographic Ambiguity

Different words with the same spelling might have completely different pronunciations and connotations depending on the diacritical symbols and punctuation used. These indicators of how words are spoken are referred to as tashkl or harakt in Arabic, and they are written in that language. The vast majority of Arabic text is written without these signs, making it difficult to understand what is being said. For example, the Arabic word" شْعرٌ ","شَعْرٌ" means hair," شَعَرَ" means he felt, and" شِعْرٌ" means poetry [14, 15]. Below shows an example:

- Word 'علم'
- Flag/ alam/ عَلَمٌ
- Science/ ilm/ عَلَّمَ
- He taught / allama/ عَلَّمَ
- Known / ulima/ عُلِمَ

Another possible source of ambiguity in Arabic is the surrounding context. For instance, the Arabic term "قتل /killing" which means "killing" may be used in both violent and peaceful contexts. One

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 4700-4705, 2024
DOI: 10.55214/25768484.v8i6.3018

example of this usage is seen in the line "تستطيع قتل الأزهار ولكن لا تستطيع أن تمنع قدوم الربيع" which translates to "You can kill the blossoms, but you can't stop spring from arriving."

2.2.3. Arabic Dialects Vary

Although Modern Standard Arabic (MSA) is the official language of the Arabic world, most people speak a dialect of Arabic. Arabic may be understood in a variety of ways depending on where you are in the globe, or even within the same nation depending on which region you are in [16]. Below is an example of how the Arabic word for cat, which in MSA is written as Qitah/ قطة, may sound different depending on the particular Arabic dialect being spoken.

- Egyptain: Otta -أطة
- Levantine: Bisse -بسة
- Gulf: Qatwa -قطوة
- Iraqi: Bazzuna -بزونة
- Yamani: Sanoura/ demah -سنورة/دمة

In more than one dialect, some words may be spelled and spoken in the same way, but their meanings may be quite distinct from one another. For instance, the term "Nasih" in Levantine implies "overweight," whereas in Egyptian it means "clever," and in the Gulf it means "advisor." The variety of Arabic text produced by users is greater than the variety of Arabic dialects. For instance, the vast majority of Arabic material seen on social media platforms is composed in an informal vernacular. This indicates that it makes use of local terminology as well as slang words that vary depending on the region of the Arabic-speaking world in which one is located [12].

Here is a list of the different kinds of dialectal differences: (1) Classical Arabic, such as the language used in the Quran and historical texts, (2) Modern Standard Arabic, like Official language, Language of news and media, Standard writing and grammar, and The National Language, (3) Dialectal Arabic is spoken most of the time, there is no official standardization, it is the mother tongue, and it is used more and more on social media. (4) Rough classifications, such as Gulf Arabic, Levantine Arabic, Egyptian Arabic, and Maghrebi Arabic. (5) Classification Problems, such as in countries with different dialects and sub-dialects, such as Urban, Rural, and Bedouin.

2.2.4. How the Alphabets of Urdu and Farsi are Utilized

Pronunciation of the Arabic alphabet varies across different Arabic dialects. For this reason, the phonetic sound of dialectal words is often described using letters from other languages than Arabic. The Urdu and Farsi scripts are widely used here. Observe how the letter "/kaa'" in Arabic is replaced by the letter "/Qa'" in Farsi and Urdu in the image below, which is an example of a tweet written in Iraqi Arabic [17].

2.2.5. Morphological Richness

Arabic has a complex inflectional system. This is due to a complex inflectional morphology that represents gender, number, person, aspect, mood, case, state, and voice, as well as a huge variety of clitics such conjunctions, negative particles, future particles, and so on. As a consequence, it's not unusual to come across single Arabic words that have a five-word English equivalent: وَسَيدرُسُونَهَا wa+sa+ya-drus-uuna+ha "and they will study it" [17]. Arabic verbs, for example, have 5,400 inflected forms. English verbs have six, whereas Chinese verbs have one [18].

Unlike several other languages, Arabic distinguishes between a feminine and masculine gender for nouns, for example female cats are referred to as "Qetta" while male cats are called "Qett". Additionally, nouns, verbs, pronouns, and adjectives in Arabic may be either solitary, dual, or multiple. The phrase "Qettatan/ قطتان" is used to refer to two female cats, "Qettan/ قطان" is used to refer to two male cats, and the word "Qettat /قططا" is used to refer to several cats [17].

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 4700-4705, 2024
DOI: 10.55214/25768484.v8i6.3018

2.2.6. Orthographic Inconsistency

Both standard and dialectal Arabic suffer from widespread spelling inconsistencies, particularly online. While there are efforts to develop official spelling standards for dialectal Arabic through computational means, such as the work on CODA, or Conventional Orthography for Dialectal Arabic, the fact remains that one-third of all words in online comments written in Modern Standard Arabic (MSA) contain misspelled words. In addition to the Arabic alphabet, alternative writing systems, such as the Romanized variant called Arabizi, which aims to reflect the pronunciation of the words, may be found online as well [19, 17].

2.2.7. Resource Poverty (Data & Tools)

The lack of data is the primary limitation of natural language processing (NLP) techniques, including rule-based methods that need dictionaries and meticulously crafted rules, and machine learning (ML) methods that require corpora and annotated corpora. Arabic morphological analyzers and lexicons, as well as annotated and parallel data in non-news genres and in dialects, are few, despite the abundance of unannotated text corpora [20, 21].

2.2.8. Other Challenges with Arabic

Dots, which are often utilized in Arabic, provide another layer of intricacy. Many letters have similar or even identical structures, therefore letters are distinguished by the number of dots and their placements, so the letters (n-ن, b-ب, t-ت) all have the same structure but with distinct dot locations and numbers [23].

In addition, Arabic lacks the use of capital letters, a feature crucial to the recognition of nouns in natural language processing, particularly Named Entity Recognition [24, 25, 23].

Stemming is an important step in natural language processing in Arabic. Stemming generates verb stems or roots by extracting affixes or linguistic rules from a lexicon to recognize verb patterns and extract the root. In Arabic, there are two primary stemming techniques: root-based stemming and light stemming. There is no method for testing roots against dictionaries of any type in stemming. In the absence of a lexicon, the extracted roots may be invalid or a jumbled collection of letters. Until date, there has been no optimal method for indexing Arabic words [27].

Arabic also has several additional morphological quirks. By adding the definite article "the" (الـ), an uncertain word becomes a definite one.

3. Conclusion

Aside from apparent concerns like resource scarcity, Arabic presents a variety of hurdles to NLP, including orthographic ambiguity, morphological richness, dialectal variances, and orthographic noise. While they are not necessarily unique to Arabic, their combination makes Arabic processing especially difficult. Arabic natural language processing (NLP) faces a great deal of difficulty, but over the last four decades, it has also achieved a great deal of success and made significant advancements. The fact that it has been steadily progressing in a favorable direction, if sometimes slowly, gives us reason for optimism. We anticipate that the market for Arabic natural language processing will see significant expansion over the next decade or two. This is in line with increased needs and expectations for language technology around the globe, as well as a growth in natural language processing (NLP) research and development in the Arab world. The Arab world is quickly becoming a highly fertile environment ripe for huge advances because of the rising number of academics and developers working on natural language processing there.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 4700-4705, 2024
DOI: 10.55214/25768484.v8i6.3018

References

[1] "Arabic Natural Language Processing," Nyu Abu Dhabi, [Online]. Available: HTTPS://NYUAD.NYU.EDU/EN/RESEARCH/FACULTY-LABS-AND-PROJECTS/COMPUTATIONAL-APPROACHES-TO-MODELING-LANGUAGE-LAB/RESEARCH/ARABIC-NATURAL-LANGUAGE-PROCESSING.HTML.

[2] "Technology Innovation Institute Announces Launch of NOOR, the World's Largest Arabic NLP Model," Abu Dhabi's Advanced Technology Research Council, [Online]. Available: https://www.tii.ae/news/technology-innovation-institute-announces-launch-noor-worlds-largest-arabic-nlp-model.

[3] O. Obeid, N. Zalmout, S. Khalifa, D. Taji, M. Oudah, B. Alhafni, F. Al-Eryani, A. Erdmann and N. Habash, "CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing," *The 12th Language Resources and Evaluation Conference (LREC 2020)*, pp. 01-10, 2020.

[4] N. Garg and K. Sharma, "Text pre-processing of multilingual for sentiment analysis based on social network data," *International Journal of Electrical and Computer Engineering (IJECE)*, pp. 776-784, 2022.

[5] M. Mohammad, S. A. Afag, E. Z. Mohammed, E. A. Rihab and E. Yasir, "Developing Two Different Novel Techniques for Arabic Text Stemming," *Intelligent Information Management*, pp. 1-23, 2019.

[6] G. Badaro, R. Baly, H. Hajj, W. El-Hajj, K. B. Shaban, N. Habash, A. Al-Sallab and A. Hamdi, "A Survey of Opinion Mining in Arabic: A Comprehensive System Perspective Covering Challenges and Advances in Tools, Resources, Models, Applications, and Visualizations," *Association for Computing Machinery*, pp. 01-48, 2019.

[7] I. Tanfouri, G. Tlik and F. Jarray, "An automatic arabic text summarization system based on genetic algorithms," *Procedia Computer Science*, pp. 195-202, 2021.

[8] M. Abdelhamid, F. Azouaou and S. Batata, "A Survey of Plagiarism Detection Systems: Case of Use with English, French and Arabic Languages," *arXiv*, pp. 01-28, 2022.

[9] N. A. Alhassan, A. Saad Albarrak, S. Bhatia and P. Agarwal, "A Novel Framework for Arabic Dialect Chatbot Using Machine Learning," *Computational Intelligence and Neuroscience*, pp. 01-11, 2022.

[10] W. Gomaa and A. Fahmy, "Arabic Short Answer Scoring with Effective Feedback for Students," *International Journal of Computer Applications*, pp. 35-41, 2014.

[11] W. Gomaa and A. Fahmy, "Short Answer Grading Using String Similarity And Corpus-Based Similarity," *International Journal of advanced Computer Science and Applications (IJACSA)*, pp. 115-120, 2012.

[12] H. Rababah and A. T. Al-Taani, "An automated scoring approach for Arabic short answers essay questions," *2017 8th International Conference on Information Technology (ICIT)* , pp. 697-702, 2017.

[13] J. Zakraoui, M. Saleh, S. Al-Maadeed and J. M. Alja'am, "Arabic Machine Translation: A Survey With Challenges and Future Directions," *IEEE Access*, pp. 161445-161468, 2021.

[14] F. Husain, "Arabic Natural Language Processing," 14 MAY 2022. [Online]. Available: https://infoscilab.ku.edu.kw/arabic-natural-language-processing/.

[15] A. Farghaly, A. Farghaly, K. Shaalan and Khaled, "Arabic Natural Language Processing: Challenges and Solutions," *ACM Transactions on Asian Language Information Processing (TALIP)*, pp. 01-22, 2009.

[16] M. a. H. N. Diab, O. Rambow, M. Altantawy and Y. Benajiba, "COLABA: Arabic dialect annotation and processing," *LREC Workshop on Semitic Language Processing*, pp. 66-74, 2010.

[17] N. Habash, F. Al-Eryani, S. Khalifa, O. Rambow, D. Abdulrahim, A. Erdmann, er, R. Faraj, W. Zaghouani, H. Bouamor, N. Zalmout, S. Hassan, F. Alshargi, S. Alkhereyf and B. Abdulkareem, "Unified Guidelines ; Resources for Arabic Dialect Orthography}," *LREC 2018*, pp. 3628-3636, 2019.

[18] J. Zahir, "IADD: An integrated Arabic dialect identification dataset," *Data in Brief*, p. 107777, 2022.

[19] F. Husain and O. Uzuner, "Investigating the Effect of Preprocessing Arabic Text on Offensive Language and Hate Speech Detection," *Association for Computing Machinery*, pp. 01-20, 2022.

[20] K. Darwish, N. Habash, M. Abbas, H. Al-Khalifa, H. T. Al-Natsheh, S. R. El-Beltagy, H. Bouamor, K. Bouzoubaa, V. Cavalli-Sforza, W. El-Hajj, M. Jarrar and H. Mubarak, "A Panoramic Survey of Natural Language Processing in the Arab World," *arXiv*, pp. 01-34, 2020.

[21] K. Ryding and D. Wilmsen, "The Cambridge Handbook of Arabic Linguistics," *Cambridge University Press*, pp. 425 - 504, 2021.

[22] N. Habash, F. Al-Eryani, S. Khalifa, O. Rambow, D. Abdulrahim, A. Erdmann, R. Faraj, W. Zaghouani, H. Bouamor, N. Zalmout, S. Hassan, F. Alshargi, S. Alkhereyf, B. Abdulkareem, R. Eskander and Sal, "Unified Guidelines and Resources for Arabic Dialect Orthography," *In Proceedings of the Language Resources and Evaluation Conference (LREC)*, pp. 3628-3637, 2019.

[23] N. Habash, "Introduction to Arabic Natural Language Processing," *Morgan & Claypool Publishers*, pp. 1-187, 2010.

[24] S. Larabi Marie-Sainte, N. Alalyani, S. Alotaibi, S. Ghouzali and I. Abunadi, "Arabic Natural Language Processing and Machine Learning-Based Systems," *IEEE Access*, pp. 7011-7020, 2019.

[25] B. D. S. Sahu and R. Vadhwani, "Web spam detection using different features," International Journal of Soft Computing and Engineering (IJSCE), p. 70–73, 2011.

[26] H. Toda and R. Kataoka, "A Search Result Clustering Method Using Informatively Named Entities," *Association for Computing Machinery*, p. 81–86, 2005.

[27] M. Alobed, A. M M Altrad, Z. Binti Abu Bakar and N. Zamin, "AUTOMATED ARABIC ESSAY SCORING BASED ON HYBRID STEMMING WITH WORDNET," *Information Retrieval and Knowledge Management*, pp. 55-67, 2021.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 4700-4705, 2024
DOI: 10.55214/25768484.v8i6.3018