DOI:10.1145/3447735

BY KAREEM DARWISH, NIZAR HABASH, MOURAD ABBAS, HEND AL-KHALIFA, HUSEEIN T. AL-NATSHEH, HOUDA BOUAMOR, KARIM BOUZOUBAA, VIOLETTA CAVALLI-SFORZA, SAMHAA R. EL-BELTAGY, WASSIM EL-HAJJ, MUSTAFA JARRAR, AND HAMDY MUBARAK

A Panoramic Survey of Natural Language Processing in the Arab World

THE TERM *NATURAL language* refers to any system of symbolic communication (spoken, signed, or written) that has evolved naturally in humans without intentional human planning and design. This distinguishes natural languages such as Arabic and Japanese from artificially constructed languages such as Esperanto or Python. Natural language processing (NLP), also called computational linguistics or human language technologies, is the subfield of artificial intelligence (AI) focused on modeling natural languages to build applications such as speech recognition and synthesis, machine translation, optical character recognition (OCR), sentiment analysis (SA), question answering, and dialogue systems. NLP is a highly interdisciplinary field with connections to computer science, linguistics, cognitive science, psychology, mathematics, and others.

Some of the earliest AI applications were in NLP (machine translation, for example); and the last decade (2010–2020) in particular has witnessed an incredible increase in quality, matched with a rise in public awareness, use, and expectations of what may have seemed like science fiction in the past. NLP researchers pride themselves on developing language-independent models and tools that can be applied to all human languages. Machine translation systems, for example, can be built for a variety of languages using the same basic mechanisms and models. However, the reality is that some languages (English and Chinese) do get more attention than others (Hindi and Swahili). Arabic, the primary language of the Arab world and the religious language of millions of non-Arab Muslims, is somewhere in the middle of this continuum. Though Arabic NLP has many challenges, it has seen many successes and developments.

Next, we discuss Arabic's main challenges as a necessary background, and we present a brief history of Arabic NLP. We then survey a number of its research areas, and close with a critical discussion of the future of Arabic NLP. An extended version of this article including almost 200 citations and links is on Arxiv.[a]

Arabic and Its Challenges

Arabic today poses a number of modeling challenges for NLP: morphological richness, orthographic ambiguity,

a https://arxiv.org/abs/2011.12631

ALL IMAGES BY ANDRIJ BORYS ASSOCIATES. USING SHUTTERSTOCK

dialectal variations, orthographic noise, and resource poverty. We do not include issues of right-to-left Arabic typography, which is an effectively solved problem (although not universally implemented).

Morphological richness. Arabic words have numerous forms resulting from a rich inflectional system that includes features for gender, number, person, aspect, mood, case, and a number of attachable clitics. As a result, it is not uncommon to find single Arabic words that translate into five-word English sentences: وَسَيَدْرُسُونَهَا wa+sa+ya-drus-uuna+ha 'and they will study it.' This challenge leads to a higher number of unique vocabulary types compared to English, which is challenging for machine learning models.

Orthographic ambiguity. The Arabic script uses optional diacritical marks to represent short vowels and other phonological information that is important to distinguish words from each other. These marks are almost never used outside of religious texts and children's literature, which leads to a high degree of ambiguity. Educated Arabs do not usually have a problem with reading undiacritized Arabic, but it is a challenge for Arabic learners and computers. This out-of-context ambiguity in Standard Arabic leads to a staggering 12 analyses per word on average: for example, the readings of the word كتبت ktbt (no diacritics) includes كَتَبْتُ katabtu 'I wrote,' كَتَبَتْ katabat 'she wrote,' and the quite semantically distant كَتَبِتْ ka+t~ibit 'such as Tibet.'

Dialectal variation. Arabic is also not a single language but rather a family of historically linked varieties, among which Standard Arabic is the official language of governance, education, and the media, while the other varieties, so-called dialects, are the languages of daily use in spoken, and increasingly written, form. Arab children grow up learning their native dialects, such as Egyptian, Levantine, Gulf, or Moroccan Arabic, which have their own grammars and lexicons that differ from each other and from Standard Arabic. For example, the word for 'car' is سيارة syArp (sayyaara) in Standard Arabic, عربية Erbyp (arabiyya) in Egyptian Arabic, كرهبة krhbp (karhba) in Tunisian Arabic, and موتر mwtr (motar) in Gulf Arabic. The differences can be significant to the point that using Standard Arabic tools on dialectal Arabic leads to quite sub-optimal performance.

Orthographic inconsistency. Standard and dialectal Arabic are both written with a high degree of spelling inconsistency, especially on social media: A third of all words in Modern Standard Arabic (MSA) comments online have spelling errors; and dialectal Arabic has no official spelling standards, although there are efforts to develop such standards computationally, such as the work on CODA, or Conventional Orthography for Dialectal Arabic. Furthermore, Arabic can be encountered online written in other scripts, most notably, a Romanized version called Arabizi that attempts to capture the phonology of the words.

Resource poverty. Data is the bottleneck of NLP; this is true for rule-based approaches that need lexicons and carefully created rules, and for machine learning (ML) approaches that need corpora and annotated corpora. Although Arabic unannotated text corpora are quite plentiful, Arabic morphological analyzers and lexicons as well as annotated and parallel data in non-news genre and in dialects are limited.

None of the issues mentioned here are unique to Arabic—for example, Turkish and Finnish are morphologically rich; Hebrew is orthographically ambiguous; and many languages have dialectal variants. However, the combination and degree of these phenomena in Arabic creates a particularly challenging situation for NLP research and development. Additional information has been published on Arabic computational processing challenges.[4,5]

A Brief History of NLP in the Arab World

Historically, Arabic NLP can be said to have gone through three waves. The first wave was in the early 1980s with the introduction of Microsoft MS-DOS 3.3 with Arabic language support. In 1985, the first Arabic morphological analyzer was developed by Sakhr Software. Most of the research in that period focused on morphological analysis of Arabic text by using rule-based approaches. Sakhr has also continued leading research and development in Arabic computational linguistics by developing the first syntactic and semantic analyzer in 1992 and Arabic optical character recognition in 1995. Sakhr also produced many commercial products and solutions, including Arabic-to-English machine translation, Arabic text-to-speech, and an Arabic search engine. This period almost exclusively focused on Standard Arabic with a few exceptions related to work on speech recognition.

The second wave was during the

years 2000-2010. Arabic NLP gained increasing importance in the Western world especially after September 11. The U.S. funded large projects for companies and research centers to develop NLP tools for Arabic and its dialects, including machine translation, speech synthesis and recognition, information retrieval and extraction, text-to-speech, and named entity recognition. Most of the systems developed in that period used machine learning, which was on the rise in the field of NLP as a whole. In principle, ML required far less linguistic knowledge than rule-based approaches and was fast and more accurate. However, it needed a lot of data, some of which was not easy to collect, for example, dialectal Arabic to English parallel texts. Arabic's rich morphology exacerbated the data dependence further. So, this period saw some successful instances of hybrid systems that combine rule-based morphological analyzers with ML disambiguation which relied on the then newly created Penn Arabic Treebank (PATB). The leading universities, companies, and consortia at the time were Columbia University, the University of Maryland, IBM, BBN, SRI, the Linguistic Data Consortium (LDC), and the European Language Resources Association (ELRA).

The third wave started in 2010, when the research focused on Arabic NLP came back to the Arab world. This period witnessed a proliferation of Arab researchers and graduate students interested in Arabic NLP and an increase in publications in top conferences from the Arab world. Active universities include New York University Abu Dhabi (NYUAD),[b] American University in Beirut (AUB), Carnegie Mellon University in Qatar (CMUQ), King Saud University (KSU), Birzeit University (BZU), Cairo University, and others. Active research centers include Qatar Computing Research Institute (QCRI),[c] King Abdulaziz City for Science and Technology (KACST), and more. It should be noted that there are many actively contributing researchers in smaller groups across the Arab world. This period also overlapped with two major independent developments: the rise of deep learning and neural models, and the rise of social media. The first development affected the direction of research, pushing it further into the ML space; the second led to the increase in social media data, which introduced many new challenges at a larger scale: more dialects and more noise. This period also witnessed a welcome increase in Arabic language resources and processing tools, and a heightened awareness of the importance of AI for the future of the region—for example, the UAE now has a ministry specifically for AI. Finally, new young and ambitious companies such as Mawdoo3 are competing for a growing market and expectations in the Arab world.

The success of word embedding models trained on unannotated data and resulting in improved performance for NLP tasks with little or no feature engineering has led to many contributions in Arabic NLP.

Arabic Tools and Resources

We organize this section on Arabic tools and resources into two parts: first, we discuss enabling technologies which are the basic resources and utilities that are not user-facing products; and second, we discuss a number of advanced user-targeting applications.

Resource construction is a lengthy and costly task that requires significant teamwork among linguists, lexicographers, and publishers over an extended period of time.

Corpora. NLP relies heavily on the existence of corpora for developing and evaluating its models, and the performance of NLP applications directly depends on the quality of these corpora. Textual corpora are classified at a high level as annotated and unannotated corpora.

Annotated corpora are a subset of unannotated corpora that have been enriched with additional information such as contextual morphological analyses, lemmas, diacritizations, part-of-speech tags, syntactic analyses, dialect IDs, named entities, sentiment, and even parallel translations. The more information, the costlier the process is to create such corpora. For Arabic, the main collections of annotated corpora were created in its second wave, mostly outside the Arab world. The most notable annotated resource is the LDC's Penn Arabic Treebank (PATB), which provides

b http://www.camel-lab.com
c https://alt.qcri.org/

Among the challenges facing Arabic NER is the lack of letter casing, which strongly helps English NER, and the high degree of ambiguity, including especially confusable proper names and adjectives.

a relatively large MSA corpus that is morphologically analyzed, segmented, lemmatized, tagged with fine-grained parts of speech, diacritized, and parsed. PATB has enabled much of the Arabic NLP research since its creation. The Prague Arabic Dependency Treebank (PADT) was the first dependency representation treebank for Arabic. The Columbia Arabic Treebank (CATiB) was an effort to develop a simplified dependency representation with a faster annotation scheme for MSA. The University of Leeds' Quranic Arabic Corpus is a beautifully constructed treebank that uses traditional morpho-syntactic analyses of the Holy Quran. With the rising interest in dialectal data, there have been many efforts to collect and annotate dialectal data. The LDC was first to create a Levantine and an Egyptian Arabic Treebank.

In the Arab world, the efforts are relatively limited in terms of creating annotated corpora. Examples include BZU's Curras, the Palestinian Arabic annotated corpus, NYUAD's Gumar, the Emirati Arabic annotated corpus, and Al-Mus'haf Quranic Arabic corpus. Another annotation effort with a focus on MSA spelling and grammar correction is the Qatar Arabic Language Bank (QALB), a project involving Columbia and CMUQ. Other specialized annotated corpora developed in the Arab world include NYUAD's parallel gender corpus with sentences in masculine and feminine for anti-gender bias research, the Arab-Acquis corpus pairing Arabic with all of Europe's languages for a portion of European parliamentary proceedings, and the MADAR corpus of parallel dialects created in collaboration with CMUQ.

In contrast to annotated corpora, there are many unannotated datasets. Most large datasets also started outside the Arab world, such as the Agence France Presse document collection, which is heavily used for Arabic information retrieval evaluation, the LDC's Arabic Gigaword, Arabic Wikipedia, and the ArTenTen corpus. Important collections in the Arab World include: the International Corpus of Arabic of Bibliotheca Alexandrina; Shamela, a large-scale corpus (1B words) covering the past 14 centuries of Arabic; the Tashkeela corpus, containing 75M fully vocalized words (National Computer Science Engineering School in Algeria); NYUAD's Gumar Gulf Arabic corpus, containing over 100M words of Internet novels; and Abu El-Khair corpus (Umm Al-Qura University, Saudi Arabia). The success of word embedding models trained on unannotated data and resulting in improved performance for NLP tasks with little or no feature engineering has led to many contributions in Arabic NLP. The more recent appearance of contextualized embeddings trained on unannotated data, such as BERT, is creating promising possibilities for improving many Arabic NLP tasks. At the time of writing this article, a handful of contextualized embedding models are known to support Arabic including Multilingual BERT, AraBERT (AUB), GigaBert (Ohio State University), Marbert (University of British Columbia), and QARiB (QCRI).

Lexical resources. We can distinguish three types of lexical resources (that is, lexicons, dictionaries, and databases): morphological resources that encode all inflected forms of words; lexical resources that are lemma based, such as machine-readable monolingual and multilingual dictionaries; and semantic resources that link lemmas to each other, such as wordnets and ontologies. These resources are useful for a variety of NLP tasks.

Some of the earliest publicly available Arabic lexical resources were created outside of the Arab world in the second wave mentioned earlier. The Buckwalter Arabic Morphological Analyzer (BAMA), with its extended version called Standard Arabic Morphological Analyzer (SAMA), both available from the LDC, provided one of the first stem databases with tags and morphological solutions, and are used in a number of tools. Elixir-FM is a functional morphology analyzer developed at Charles University in the Czech Republic. The DIINAR Arabic morphological database is a full form resource developed in France. The Tharwa lemma-based lexicon was developed at Columbia University and included 70k entries in Egyptian Arabic, MSA, and English; it was later extended with Levantine Arabic.

Arabic WordNet is a semantic lexicon consisting of about 11k synsets, with subset and superset relationships between concepts, and linked to a number of other languages through the Global WordNet effort. This effort was done by a number of American and European universities. And the Arabic VerbNet classifies verbs that have the same syntactic descriptions and argument structure (University of Konstanz, Germany).

Some of the efforts in the Arab world led to multiple notable resources. Al-Khalil analyzer is a large morphological database for Arabic developed by researchers in Morocco and Qatar. Calima Star is an extension of the BAMA/SAMA family done at NYUAD and is part of the CAMeL Tools toolkit. BZU developed a large Arabic lexicographic database constructed from 150 lexicons that are diacritized and standardized. The MADAR project (NYUAD and CMUQ) includes a lexicon with 47k lemma entries of parallel statements in 25 city dialects. Other lexicons have been developed for Algerian, Tunisian, and Moroccan. Finally, in terms of semantic lexical resources, the BZU Arabic Ontology is a formal Arabic wordnet with more than 20k concepts that was built with ontological analysis in mind and is linked to the Arabic Lexicographic Database, Wikidata, and other resources.

More Arabic resources can be found in known international repositories (namely ELRA/ELDA, LDC, and CLARIN) or directly from their authors' websites.[4,5,10] Unfortunately, many are not interoperable, have been built using different tools and assumptions, released under propriety licenses, and few are comprehensive. Serious, well-planned, and well-coordinated investment in resources will be instrumental for the future of Arabic NLP.

Morphological processing. Given the challenges of Arabic morphological richness and ambiguity, morphological processing has received a lot of attention. The task of morphological analysis refers to the generation of all possible readings of a particular undiacritized word out of context. Morphological disambiguation is about identifying the correct in-

context reading. This broad definition allows us to think of word-level tasks such as part-of-speech (POS) tagging, stemming, diacritization, and tokenization as sub-types of morphological disambiguation that focus on specific aspects of ambiguity.

Most work on Arabic morphological analysis and disambiguation is on MSA; however, there is a growing number of efforts on dialectal Arabic. There are a number of commonly used morphological analyzers for Standard and dialectal Arabic (Egyptian and Gulf), including BAMA, SAMA, Elixir-FM, Al-Khalil, CALIMA Egyptian, and CALIMA Star. Some of the morphological disambiguation systems disambiguate the analyses that are produced by a morphological analyzer using PATB as a training corpus, for example, MADAMIRA (initially developed at Columbia University) and other variants of it from NYUAD. Farasa (from QCRI) uses independent models for tokenization and POS tagging.

Syntactic processing. Syntactic parsing is the process of generating a parse tree representation for a sentence that indicates the relationship among its words. For example, a syntactic parse of the sentence أَرَقَ ٱلطَّالِبُ ٱلْجَدِيدُ ٱلْكِتَابَ ٱلْجَدِيدَ '[lit.] *read the-student the-book the-new*; 'the student read the new book,' would indicate that the adjective *the-new* modifies the noun *the-book*, which itself is the direct object of the verb *read*.

There are many syntactic representations. Most commonly used in Arabic are the PATB constituency representation, the CATiB dependency representation, and the Universal Dependency (UD) representation. All of these were developed outside of the Arab world. The UD representation is an international effort, where NYUAD is the representative of the Arab world on Arabic.

The most popular syntactic parsers for Arabic are: Stanford, Farasa (QCRI), and CamelParser (NYUAD). Stanford is a statistical parser from the Stanford Natural Language Processing Group that can parse English, German, Arabic, and Chinese. For Arabic, it uses a probabilistic context free grammar that was developed based on PATB. Farasa is an Arabic NLP toolkit that provides syntactic constituency and dependency parsing. CamelParser is a dependency parser trained on CATiB treebank using MaltParser, a language-independent and data-driven dependency parser. A discussion and survey of some of the Arabic parsing work is presented in Habash.[5]

Named entity recognition (NER) is the task of identifying one or more consecutive words in text that refer to objects that exist in the real world (named entities), such as organizations, persons, locations, brands, products, foods, and so forth. NER is essential for extracting structured data from an unstructured text, relationship extraction, ontology

population, classification, machine translation, question answering, and other applications. Among the challenges facing Arabic NER compared to English NER is the lack of letter casing, which strongly helps English NER, and the high degree of ambiguity, including especially confusable proper names and adjectives, for example, كريم *kariym* can be the name 'Kareem' or the adjective 'generous.'

Arabic NER approaches include the use of hand-crafted heuristics, machine learning, and hybrids of both with heavy reliance on gazetteers. Much of the earlier work on Arabic NER focused on formal text, typically written in MSA. However, applying models trained on MSA text to social media (mostly dialectal) text has led to unsatisfactory results. Recent contextualized embeddings and other deep learning approaches such as sequence-to-sequence models and convolutional neural networks have led to improved results for Arabic NER. As with other utilities, early research was done outside of the Arab world, but more work is now happening in the Arab world. An extensive list of Arabic NER challenges and solutions can be found in Shaalan.[9]

Dialect identification (DID) is the task of automatically identifying the dialect of a particular segment of speech or text of any size: word, sentence, or document. This task has been attracting increasing attention in NLP for a number of language varieties. DID has been shown to be important for several NLP tasks where prior knowledge about the dialect of an input text can be helpful, such as machine translation, sentiment analysis, and author profiling.

Early Arabic multi-dialectal data sets and models focused on the regional level. The Multi Arabic Dialects Application and Resources (MADAR) project aimed to create a finer grained dialectal corpus and lexicon. The data was used for dialectal identification at the city level of 25 Arab cities, and was used in a shared task for DID. The main issue with that data is that it was commissioned and not naturally occurring. Concurrently, larger Twitter-based datasets covering 10-to-21 countries were also introduced. The Nuanced Arabic Dialect Identification (NADI) shared task followed earlier pioneering works by providing country-level dialect data for 21 Arab countries, and introduced a province-level identification task aiming at exploring a total of 100 provinces across these countries. Earlier efforts started in the west, most notably in Johns Hopkins University, but more work is happening now in the Arab world, at NYUAD and QCRI, for example.

Infrastructure. To aid the development of NLP systems, a number of multi-lingual infrastructure toolkits have been developed, such as GATE,[d] Stanford CoreNLP,[e] and UIMA.[f] They offer researchers easy access to several tools through command-line interfaces (CLIs) and application programming interfaces (APIs), thus eliminating the need to develop them from scratch every time. While Arabic NLP has made significant progress with the development of several enabling tools, such as POS taggers, morphological analyzers, text classifiers, and syntactic parsers, there is a limited number of homogeneous and flexible Arabic infrastructure toolkits that gather these components. MADAMIRA is a Java-based system providing solutions to fundamental NLP tasks for Standard and Egyptian Arabic. These tasks include diacritization, lemmatization, morphological analysis and disambiguation, POS tagging, stemming, glossing, (configurable) tokenization, base-phrase chunking, and NER.[g] Farasa[h] is a collection of Java libraries and CLIs for MSA. These include separate tools for diacritization, segmentation, POS tagging, parsing, and NER. SAFAR[i] is a Java-based framework bringing together all layers of Arabic NLP: resources, pre-processing, morphology, syntax, and semantics. CAMeL Tools is a recently developed collection of open source tools, developed in Python, that supports both MSA and Arabic dialects.[j] It currently provides APIs and CLIs for pre-processing, morphological modeling, dialect identification, NER, and sentiment analysis. Other notable efforts include AraNLP, ArabiTools,[k] and Adawat.[l] A feature comparison of some Arabic infrastructures can be found in Obeid,[8] while a detailed survey and a software engineering comparative study can be found in Jaafar.[6]

Arabic NLP Applications

Machine translation (MT) is one of the earliest and most worked on areas in NLP. The task is to map input text in a source language such as English

d https://gate.ac.uk

e https://stanfordnlp.github.io/CoreNLP/
f https://uima.apache.org/d/uimaj-current/
g https://camel.abudhabi.nyu.edu/madamira/
h https://farasa.qcri.org/
i http://arabic.emi.ac.ma/safar/
j https://github.com/CAMeL-Lab
k https://www.arabitools.com/
l http://adawat.sourceforge.net/

to an output text in a target language such as Arabic. Early MT research was heavily rule-based; however, now it is almost completely corpus-based using a range of statistical and deep learning models, depending on resource availability.

For MSA, parallel data in the news domain is plentiful.[m] There are other large Arabic parallel collections under the OPUS project and as part of the United Nations' six-language parallel corpus. Other specialized corpora include the Arab-Acquis corpus pairing with European languages developed in NYUAD, and the AMARA educational domain parallel corpus developed by QCRI. Dialectal parallel data are harder to come by and most are commissioned translations.

There are many other efforts in Statistical MT (SMT) from and to Arabic. Recently, deep neural networks have been adopted for Arabic machine translation. While most researched MT systems for Arabic target English, there have been efforts on MT for Arabic and other languages, including Chinese, Russian, Japanese, and all of the European Union languages.

MT for Arabic dialects is more difficult due to limited resources, but there are noteworthy efforts exploiting similarities between MSA and dialects in universities and research group around the world. Finally, there is a notable effort on Arabic sign-language translation at King Fahd University of Petroleum and Minerals. For recent surveys of Arabic MT, see Ameur.[1] Despite all these contributions, much research work is still needed to improve the performance of machine translation for Arabic.

Pedagogical applications (PA) focus on building tools to develop or model for four major skills: reading, writing, listening, and speaking. Arabic PA research has solely focused on MSA. PA systems can be distinguished in terms of their target learners as first language (L1) or second (foreign) language (L2) systems. This distinction can be problematic since, for Arabs, learning to read MSA is somewhat akin to reading a foreign tongue due to its lexical and syntactic divergence from native dialects. We focus our Arabic PA discussion on (a) computer-assisted language learning (CALL) systems, (b) readability assessment, and (c) resource-building efforts.

CALL systems utilize NLP enabling technologies to assist language learners. There has been a number of efforts in Arabic CALL exploring a range of resources and techniques. Examples include the use of Arabic grammar and linguistic analysis rules to help learners identify and correct a variety of errors; and multi-agent tutoring systems that simulate the instructor, the student, the learning strategy, and include a logbook to monitor progress, and a learning interface. Another approach focuses on enriching the reading experience with concordances, text-to-speech, morpho-syntactic analysis, and auto-generated quiz questions.

Readability assessment is the task of automatic identification of a text's readability, that is, its ability to be read and understood by its reader employing an acceptable amount of time and effort. There has been a range of approaches for Arabic L1 and L2 readability. On one end, we find formulas using language-independent variables such as text length, average word length, and average sentence length, number of syllables in words, the relative rarity or absence of dialectal alternatives, and the presence of less common letters. Others integrate Arabic morphological, lexical, and syntactic features with supervised machine learning approaches.

Although some progress has been made for both L1 and L2 PA, the dearth of resources compared with English remains the bottleneck for future progress. Resource-building efforts have focused on L1 readers with particular emphasis on grade school curricula. There is a push to inform the enhancement of curricula using pedagogical tools and to compare curricula across Arab countries. The L2 PAs are even more constrained, with limited corpora and a disproportionate focus on beginners.[n] There is a definite need

Current NLP methods for Arabic language dialogue are mostly based on handcrafted rule-based systems and methods that use feature engineering.

m Linguistic Data Consortium (LDC) resources: LDC2004T18, LDC2004T14, and LDC2007T08.

n https://learning.aljazeera.net/en

It is time to have an association for Arabic language technologists that brings together talent and resources and sets standards for the Arabic NLP community.

for augmenting these corpora in a reasoned way, taking into consideration different text features and learners, both young and old, beefing up the sparsely populated levels with authentic material, and exploiting technologies such as text simplification and text error analysis and correction. Learner corpora, which as the name suggests are produced by learners of Arabic, can inform the creation of tools and corpora. A recent effort developed a large-scale Arabic readability lexicon compatible with an existing morphological analysis system.

Information retrieval and question answering. With the increasing volume of Arabic content, information retrieval, or search, has become a necessity for many domains, such as medical records, digital libraries, web content, and news. The main research interests have focused on retrieval of formal language, mostly in the news domain, with ad hoc retrieval, OCR document retrieval, and cross-language retrieval. The literature on other aspects of retrieval continues to be sparse or non-existent, though some of these aspects have been investigated by industry. Other aspects of Arabic retrieval that have received some attention include document image retrieval, speech search, social media and web search, and filtering.[3] However, efforts on different aspects of Arabic retrieval continue to be deficient and severely lag behind efforts in other languages. Examples of unexplored problems include searching Wikipedia, which contains semi-structured content, religious texts, which often contain semi-structured data such as chains of narrations, rulings, and commentaries, Arabic forums, which are very popular in the Arab world and constitute a significant portion of the Arabic Web, and poetry. To properly develop algorithms and methods to retrieve such content, standard test sets and clear usage scenarios are required. We expect that recent improvements in contextual embeddings can positively impact the effectiveness of many retrieval tasks.

Another information retrieval-related problem is question answering, which comes in many flavors, the most common of which is attempting to identify a passage or a sentence that answers a question. Performing such a task may employ a large set of NLP tools such as parsing, NER, coreference resolution, and text semantic representation. There has been limited research on this problem, and existing commercial solutions such as Ujeeb.com are rudimentary.

Dialogue Systems

Automated dialog systems capable of sustaining a smooth and natural conversation with users have attracted considerable interest from both research and industry in the past few years. This technology is changing how companies engage with their customers among many other applications. While commercial dialog systems by big multinational companies such as Amazon's Alexa, Google's Home, and Apple's Siri support many languages, only Apple's Siri supports Arabic with limited performance. There are some strong recent competitors in the Arab world, particularly Arabot[o] and Mawdoo3's Salma.[p]

While there is an important growing body of research on English language dialog systems, current NLP methods for Arabic language dialogue are mostly based on handcrafted rule-based systems and methods that use feature engineering. Among the earliest research efforts on Arabic dialog applications is the Quran chatbot, where the conversation length is short since the system answers a user input with a single response. It uses a retrieval-based model as the dataset is limited by the content of the Quran. A recent approach used deep learning techniques for text classification and NER to build a natural language understanding module—the core component of any dialogue system—for the domain of home automation in Arabic. A unique dialogue system from NYUAD explored bilingual interfaces where Arabic speech can be used as input to an English bot that displays Arabic subtitles. Other works have focused on developing dialog systems for the case of Arabic dialects, as with the publicly avail-

o https://arabot.io/
p http://salma.ai/

able NYUAD Egyptian dialect chatbot *Botta*, and KSU's Saudi dialect information technology-focused chatbot *Nabiha*.

Sentiment and Emotion Analysis

Sentiment analysis (SA), or opinion mining, is the task of identifying the affective states and subjective information in a text. For example, an Egyptian Arabic movie review such as أحسن فيلم السنة دي! 'the best movie this year!' is said to indicate a positive sentiment. SA is a very powerful tool for tracking customer satisfaction, carrying out competition analysis, and generally gauging public opinion towards a specific issue, topic, or product. SA has attracted a lot of attention in the Arabic research community during the last decade, connected with the availability of large volumes of opinionated and sentiment reflecting data from Arabic social media. Early Arabic SA efforts focused on the creation of needed resources such as sentiment lexicons, training datasets, and sentiment treebanks, as well as shared task benchmarks. Arabic SA solutions span a range of methods from the now conventional use of rules and lexicons to machine learning based methods as well as hybrid approaches employing morphological and syntactic features. Recently, fine-tuning large pre-trained language models has achieved improved Arabic SA results. Arabic emotion detection is a closely related topic that has attracted some attention recently. It aims to identify a variety of emotions in text such as anger, disgust, surprise, and joy. Similar to how SA resources and models started maturing, a lot of work still needs to be done in emotion detection. Another related problem is stance detection, which attempts to identify positions expressed on a topic or towards an entity. Stances are often expressed using non-sentiment words. For a recent comprehensive survey on the status of Arabic SA and the future directions, see Badaro et al.[2]

Content Moderation on Social Media

The task of content moderation is about the enforcement of online outlets' policies against posting user comments that contain offensive language, hate speech, cyber-bullying, and spam, among other types of inappropriate or dangerous content.[q] Such content cannot be easily detected given the huge volume of posts, dialectal variations, creative spelling on social media, and the scarcity of available data and detection tools. This area is relatively new for Arabic. One of the more active areas has to do with the detection of offensive language, which covers targeted attacks, vulgar and pornographic language, and hate speech. Initial work was performed on comments from a news site and a limited number of tweets and YouTube comments. Some works focused on adult content and others on hate speech. Recent benchmarking shared tasks included the automatic detection of such language on Twitter. Work on spam detection on Twitter is nascent and much work is required.

Future Outlook

Arabic NLP has many challenges, but it has also seen many successes and developments over the last 40 years. We are optimistic by its continuously positive albeit sometimes slow development trajectory. For the next decade or two, we expect a large growth in the Arabic NLP market. This is consistent with global rising demands and expectations for language technologies and the increase in NLP research and development in the Arab world. The growing number of researchers and developers working on NLP in the Arab world makes it a very fertile ground ready for major breakthroughs. To support this vision, we believe it is time to have an association for Arabic language technologists that brings together talent and resources and sets standards for the Arabic NLP community. Such an organization can support NLP education in the Arab world, serve as a hub for resources, and advocate for educators and researchers in changing old-fashioned university policies regarding journal-focused evaluation, and encouraging collaborations within the Arab world by connecting academic, industry, and governmental stakeholders. We also recommend more open source tools and public data be made available to create a basic development framework that lowers the threshold for joining the community, thus attracting more talent that will form the base of the next generation of Arabic NLP researchers, developers, and entrepreneurs. C

q https://www.bbc.co.uk/usingthebbc/terms/what-are-the-rules-for-commenting/

References

1. Ameur, M.S.H., Meziane, F., Guessoum, A. Arabic machine translation: A survey of the latest trends and challenges. *Computer Science Rev. 38* (2020), 100305.
2. Badaro, G.,et al. A survey of opinion mining in Arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations. *ACM Trans.Asian and Low-Resource Language Information Processing* (*TALLIP*) 18, 3 (2019), 1–52.
3. Darwish, K., Magdy, W. Arabic information retrieval. *Foundations and Trends in Information Retrieval 7*, 4 (2014), 239–342.
4. Farghaly, A., Shaalan, K. Arabic natural language processing: Challenges and solutions. *ACM Trans. Asian Language Information Processing 8*, 4 (2009), 1–22.
5. Habash, N.Y. *Introduction to Arabic Natural Language Processing, Vol. 3*. Morgan & Claypool Publishers, 2010.
6. Jaafar, Y., Bouzoubaa, K. A survey and comparative study of Arabic NLP architectures. *Intelligent Natural Language Processing: Trends and Applications*. Springer, 2018, 585–610.
7. Mubarak, H., et al. Overview of OSACT4 Arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, May 2020. European Language Resource Association, 48–52.
8. Obeid, O., et al. CAMeL tools: An open source Python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conf.*, May 2020, European Language Resources Association, 7022–7032.
9. Shaalan, K. A survey of Arabic named entity recognition and classification. *Computational Linguistics 40* (2014), 469–510.
10. Zaghouani, W. Critical survey of the freely available Arabic corpora. In *Proceedings of the Workshop on Open-Source Arabic Corpora and Processing Tools* (2014), 1–8.

Kareem Darwish, Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar.

Nizar Habash, New York University Abu Dhabi, United Arab Emirates.

Mourad Abbas, Center of Scientific and Technical Research for the Development of Arabic Language (CRSTDLA), Bouzareah, Algeria.

Hend Al-Khalifa, King Saud University, Riyadh, Saudi Arabia.

Huseein T. Al-Natsheh, Mawdoo3, Jordan.

Houda Bouamor, Carnegie Mellon University, Doha, Qatar.

Karim Bouzoubaa, Mohammed V University, Rabat, Morocco.

Violetta Cavalli-Sforza, Al Akhawayn University, Ifrane, Morocco.

Samhaa R. El-Beltagy, Newgiza University, Cairo, Egypt.

Wassim El-Hajj, American University of Beirut, Beirut, Lebanon.

Mustafa Jarrar, Birzeit University, Palestine.

Hamdy Mubarak, Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar.