18

Arabic Computational Linguistics

Nizar Habash

18.1 Introduction

Computational linguistics, or, as it is often referred to interchangeably, natural language processing or human language technologies (henceforth CL/NLP), is a large and growing highly interdisciplinary field of research that lies at the intersection of linguistics, computer science, electrical engineering, cognitive science, psychology, pedagogy, and mathematics, among other fields. The goal of CL/NLP is to develop applications that model human linguistic processes, such as speech recognition, translation, etc. These applications can serve as interfaces between humans and machines, or between humans and other humans in cross-lingual settings.

In the first two decades of the twenty-first century, research in CL/NLP led to impressive applications that have changed how humans expect machines to process language, from fast and reliable search, to almost human-quality machine translation in certain genres and languages, to speech-based personal assistants. These successes are most seen in resource-rich languages such as English, but are behind for other languages, including Arabic.

This chapter presents a brief introduction to Arabic CL/NLP, primarily targeting linguistics readers. It is organized in six sections. Section 18.2 presents a general overview of the field of CL/NLP. Section 18.3 presents the unique challenges facing Arabic CL/NLP. Section 18.4 discusses some of the common themes of CL/NLP research, and Arabic-specific solutions. Section 18.5 lists some introductory notes and resources for linguists interested in CL/NLP, and Section 18.6 offers some thoughts on the future of the field of Arabic CL/NLP.

* I would like to thank Go Inoue, Alexander Erdmann, and Nasser Zalmout for helpful discussions and feedback.

https://doi.org/10.1017/9781108277327.019 Published online by Cambridge University Press

18.2 Computational Linguistics

In this section, I present CL/NLP in terms of the types of systems it builds as well as the types of approaches it takes in building them.

18.2.1 Applications and Enabling Technologies

When discussing CL/NLP systems and research, we can distinguish between *applications* and *enabling technologies*.

Application refers to research on end-to-end systems that interface with human users. Examples include the following.

- **Information retrieval** returns documents relevant to a given query (word or a phrase).
- **Machine translation** automatically translates a document from one language to another.
- **Automatic speech recognition** automatically transcribes as text the words in an audio signal.
- **Speech synthesis** produces speech from text.
- **Optical character recognition** automatically transcribes images of words into machine-readable text.

Other prominent examples include dialogue systems, sentiment analysis, and automatic summarization.

Enabling technology refers to system components that focus on enriching or transforming text in specific ways relevant to specific CL/NLP applications, but that are not intended for the human end user. Examples include the following.

- **Tokenization** converts input words into smaller tokens that downstream components expect to work on.
- **Part-of-speech (POS) tagging** assigns each token a POS.
- **Syntactic parsing** creates a syntactic analysis of the tokens in a sentence.
- **Language/dialect identification** identifies the language or dialect of the text.

Other prominent examples include lemmatization, base-phrase chunking and named-entity recognition.

Some applications can also be *enabling* to other applications, e.g., speech-to-speech machine translation systems often employ speech recognition on the source language, text-to-text machine translation, and speech synthesis for the target language. Other enabling technologies can be considered end-user consumable in limited contexts, e.g., language identification or Arabic automatic diacritization.

https://doi.org/10.1017/9781108277327.019 Published online by Cambridge University Press

18.2.2 Approaches

Within CL/NLP, we often talk of *modelling* language. Here, *modelling* refers to designing and building computational systems (i.e., *models*) that can predict or imitate specific human linguistic competencies. In general, there are two different paradigms in CL/NLP: rule-based approaches and machine-learning approaches.

Rule-based approaches rely on human-created rules that are applied by machines to model a linguistic behaviour. These approaches, in their pure form, depend on the linguistic intuitions and competency of the rule writers. Rules can consist of large collections of lexical entries, e.g., Arabic broken plurals paired with their lemmas. Rules can also be for very specific non-lexical conditions, e.g., word diacritization reflecting the Arabic definite article's Sun/Moon letter assimilation. As with all grammars, rules are leaky and can be tedious to create.

Machine-learning (ML) approaches use algorithms to learn the 'rules' from data. Other names for these approaches include *corpus-based approaches*, in reference to the corpora of data used to train the models, as well as statistical, neural, or deep-learning approaches in specific reference to the types of algorithms used. ML approaches can be subclassified into supervised and unsupervised varieties.

Supervised approaches require data annotated with the target information to model. For instance, to train a POS-tagger, we need training sentences with the target POS sequences; to train a machine translation system, we need sentences in the source and target languages, also known as parallel corpora. And similarly, to train models of speech recognition, we need audio recordings paired with their transcripts.

Unsupervised approaches require data without additional annotation. The data can be used to model the distributional or co-occurrence properties of language, as in n-gram sequence language models (Jurafsky and Martin 2008), or word-to-vector (word2vec) models that capture the semantics of words given their contexts (Mikolov et al. 2013).

ML approaches dominate the research in CL/NLP, but they have some limitations. For ML, the size of the data set is extremely important – the more the better. The sizes of training instances range from tens of thousands into millions of tokens for reasonable performance. For many languages and tasks, this is often a challenge. In comparison with the official languages of the European Union (Koehn 2005) or the United Nations (Rafalovitch et al. 2009), most of the world's 7,000 languages and dialects are extremely poor resource-wise. ML approaches are also sensitive to the domain and genre of the training data. This is why machine translations systems that do well on news text (a very rich genre) do not fare as well in poetry or colloquial speech. When modelling resource-poor languages, rule-based approaches are a good start until enough annotations are created with real data that can benefit from ML approaches. In practice, the combination of human-created rules and machine learning is quite common, and can be seen in different parts of

https://doi.org/10.1017/9781108277327.019 Published online by Cambridge University Press

the same system. For instance, the state-of-the-art system for Arabic morphology disambiguation, MADAMIRA (Pasha et al. 2014), uses a rule-based morphological analyser and ML-based disambiguation classifiers.

A common misconception about ML approaches is that there is no human input into the modelling beyond algorithmic design. But obviously, humans are essential in creating the examples and annotations that are necessary for training the models. The quality and robustness of these models are highly dependent on the quality of the human annotations. Furthermore, all CL/NLP approaches need so-called gold references that can be used to evaluate their quality. CL/NLP evaluation is a huge area of research that is extremely important, not just for judging the quality of the final models, but also for their development.

18.3 Arabic Computational Challenges

When compared with English, Arabic poses a number of CL/NLP modelling challenges: morphological richness, orthographic ambiguity, dialectal variations, orthographic noise, and resource poverty. These issues are not necessarily unique to Arabic, but their combination makes Arabic processing particularly challenging. We will not discuss technological support challenges, such as the proper display of Arabic fonts, as these are solved problems that may not be implemented in certain frameworks for reasons varying from incompetence to lack of interest in the Arab world as a market.

18.3.1 Morphological Richness and Complexity

Arabic words have a large number of forms. This results from its rich inflectional morphology that models gender, number, person, aspect, mood, case, state, and voice, in addition to the cliticization of a number of pronouns and particles (conjunctions, prepositions, the definite article, the future particle, etc). For example, a Modern Standard Arabic (MSA) verb lemma has about 5,400 inflected forms, e.g., قال *qāl* 'to say' may realize as يقول *yaqūlu* 'he says [indicative]', فسيقولونها *fasayqūlūnahā* 'so they [masc.pl] will say it', and وقيل *waqīla* 'and it was said', among many other forms. In comparison, English has about six or so verb forms, e.g., *say/VB*, *says/VBZ*, *say/VBP*, *said/VBD*, *said/VBN*, *saying/VBG*. Arabic morphology is also complex, not just from using templatic and affixational morphological techniques, but also from (a) numerous complex morphotactics such as the *hollow* and *hamzated* verb conjugation rules and (b) the high percentage of broken plural forms, among other phenomena. These phenomena do not just make modelling more involved, but also lead to sparsity issues, since without proper abstraction features, the ML approaches cannot relate different forms of the word to each other.

https://doi.org/10.1017/9781108277327.019 Published online by Cambridge University Press

18.3.2 Ambiguity

Arabic orthography using the Arabic script employs optional diacritical marks for short vowels and consonantal gemination. The missing diacritics are not a major challenge to literate native adults. However, their absence is the main source of ambiguity in Arabic CL/NLP. In MSA, the average ambiguity is twelve analyses per word, corresponding to 2.7 lemmas (dictionary lexical entries) per word. For example, the MSA word عقد can be analysed as the following lemmas: عَقْد ʿ*aqd* 'contract', عُقْد ʿ*uqd* 'necklace', عُقَد ʿ*uqad* 'complexes' (lemma عُقْدَة ʿ*uqda*), عَقَدَ ʿ*aqada* 'to tie', or عَقَّدَ ʿ*aqqada* 'to complicate'. Some lemmas have additional analyses that vary in diacritization and features, e.g., the word for 'contract' has forms that vary in definiteness and case, as in عَقْدٌ ʿ*aqdun*, and عَقْدِ ʿ*aqdi*. The ambiguity can also result from different interpretation of clitic letters: e.g., وجد can be analysed as وَجَدَ *waǧada* 'he found' or وَجَدُ *waǧaddu* 'and the grandfather of'. This is a big challenge that necessitates models for contextual ambiguity resolution.

18.3.3 Dialects

While MSA is the main language of culture and education in the Arab world, it is no one's native tongue. Arab children grow up learning local dialects such as Egyptian, Levantine, and Gulf. Arabic dialects differ significantly from one another and from MSA in terms of their phonology, morphology, and lexicon (Watson 2007). This difference is rather severe, to the point that using MSA tools and resources for processing dialects is not sufficient. For instance, Khalifa et al. (2016) report that using a state-of-the-art tool for MSA morphological disambiguation on Gulf Arabic returns POS-tag and lemma accuracy at about 72 and 64 per cent respectively, compared to the performance on MSA, which is 96 per cent for both (Pasha et al. 2014). Dialectal morphology is particularly problematic, as despite being simpler than MSA's in certain respects (no case, reduced paradigms), it is far more complex in other respects (many more clitics). For example, the word used in the previous section has at least one additional Levantine and Egyptian dialectal reading: عَقَدّ ʿ*a*+ʾ*add* 'on the size of'. Add to the above that there is no standard orthography, an issue that leads to rampant spelling inconsistency, which we will discuss more in the next section. Finally, since the dialects and MSA coexist in a diglossic situation, developing CL/NLP systems is more challenging as multiple forms of Arabic need to be modelled together (Elfardy and Diab 2013).

18.3.4 Orthographic Inconsistency

Arabic and its dialects, as encountered online, show a huge amount of spelling inconsistency. Noise in written text is a common problem for CL/NLP when working in social media and non-edited text in any language, of course. For MSA, Zaghouani et al. (2014) report that 32 per cent of words in

https://doi.org/10.1017/9781108277327.019 Published online by Cambridge University Press

MSA comments online have spelling errors. These errors lead not only to words that are unknown by the CL/NLP systems, but also introduce unforeseen ambiguities. For example, the MSA word بادلتها can be بَادَلَتْهَا *bādalathā* 'she exchanged it', or it can be corrected as بِأَدِلَّتِهَا *biʾadillatihā* 'with her pieces of evidence'. Dialectal Arabic is more problematic since it has no standard orthography. For instance, the Egyptian Arabic word *mabiʾulhāš* 'he does not say it' has twenty-seven Arabic script spellings found with Google Search, including مبيقولهاش (≈26,000 times), ما بيقولهاش (≈13,000), مبيقلهاش (≈1,000), and مبؤلهاش (less than 10). Habash et al. (2012b) proposed a conventional orthography for dialectal Arabic (CODA), a set of guidelines for consistent spelling of Arabic dialects for CL/NLP purposes. Based on this CODA anchor, Eskander et al. (2013) reported that close to 24 per cent of Egyptian Arabic words have non-CODA-compliant spelling. In addition to the variety in Arabic script spelling, dialectal Arabic text in particular is also known to appear on social media in a non-standard romanization, often called Arabizi (Al-Badrashiny et al. 2014).

18.3.5 Resource Poverty

The existence of resources such as corpora, annotated corpora, and lexicons, is a bottleneck for research in CL/NLP . For Arabic, specific MSA genres have the lion's share of resources (Jinxi 2002; Munteanu and Marcu 2007). Most of the parallel corpora and annotated treebanks (Maamouri et al. 2004) are in the news genre, where parallel data is created 'naturally' as part of the process of news creation and dissemination across the globe. MSA has additional large sets of UN resolutions paired with other UN languages. Some commissioned MSA translations are used to benchmark performance against other language pairs (Habash et al. 2017). And a sizable effort went into Qur'ānic Arabic annotation (Dukes et al. 2013). Of course, the rich MSA news genre is important and relevant for a large number of news-oriented applications. However, applications for other genres, domains, and dialects must make do with impoverished resources, which leads to low-quality performance. A commonly encountered example is the low quality of online machine translation of tweets and social media posts. Most of the parallel data for dialect-English (Zbib et al. 2012) and dialect-MSA (Bouamor et al. 2014) is commissioned and limited.

18.4 Arabic Computational Processing Approaches and Solutions

In this section, we discuss some common themes in CL/NLP research, not necessarily specific to Arabic, but with reference to Arabic CL/NLP efforts. Then we present some details particularly in the areas of computational processing of Arabic orthography, morphology, syntax, and semantics.

https://doi.org/10.1017/9781108277327.019 Published online by Cambridge University Press

18.4.1 Common Themes in CL/NLP Research

Application-Oriented Evaluation Evaluation is central in CL/NLP research, with common references to *intrinsic* evaluations (focused on evaluating a component out of its context of use) and *extrinsic* evaluations (typically as part of an application's end-to-end use). The CL/NLP community runs many shared tasks and evaluation campaigns, usually around a specific task and with predefined training data and evaluation metrics. Metrics are themselves a big area of research; there are occasionally workshops dedicated to evaluating evaluation metrics, particularly for complex applications such as machine translation. The focus on evaluation also requires extensive quality control on the *reference* (*gold*) annotations using techniques such as inter-annotator agreement measurement.

Representation and Featurization Representation of linguistic information (words, analyses, features, etc.) is an important area of CL/NLP research, interacting heavily with questions of evaluation and application readiness. Obviously, input and output of applications must be human produceable and consumable, respectively. However, what is acceptable to humans may not be optimal for modelling. This is often handled with preprocessing and post-processing programs that manipulate the input and reformat the output. For example, when translating from English to Arabic, it may be desirable to build models on lower-case English, Alif-Hamza normalized and morphologically tokenized Arabic, and with punctuation separated from words. The English input must then be lower-cased and punctuation separated in a preprocessing step; and the Arabic output must then be detokenized, Alif-Hamza enriched and punctuation adjusted (El Kholy and Habash 2012). All of this serves to reduce data sparsity without the loss of highly informative nuances of the language. This example of text transformation is a type of featurization, where we use features to represent an input word. The term *feature* has two uses, one as in linguistics to refer to a linguistic aspect of a word, e.g., gender, number, or tense. The other use is as an ML feature used by an ML algorithm. All linguistic features can be used as ML features. But ML features can include information other than what is typically thought of as linguistic features, e.g., not so abstract lexical representations, as in spelling variants of words, or statistical information such as word probabilities. Furthermore, feature selection includes considering different granularities, e.g., the POS for a word like مُهِمَّةً *muhimmatan* 'important' can be NOM (nominal), ADJ (adjective), ADJ.FS (feminine singular adjective), or ADJ.FS.IA (indefinite accusative feminine singular adjective). ML features can also be combinations of different features and even mathematical operations on them when applicable.

Language-Independence and Cross-Lingual Adaptation Many CL/NLP techniques are language independent, and can be applied to different

https://doi.org/10.1017/9781108277327.019 Published online by Cambridge University Press

languages. Different languages may still need some different considerations, but the core algorithms tend to be the same. Additionally, many efforts consider the use of cross-lingual adaptation and information sharing to address resource poverty. Examples in Arabic include techniques to maximize the use of MSA treebanks to parse Arabic dialects (Chiang et al. 2006) or MSA–English parallel data to translate dialectal Arabic to English (Salloum and Habash 2011), and MSA–English parallel data to translate between Arabic and Hebrew (El Kholy and Habash 2015).

18.4.2 Orthographic Processing

Given the focus on applications, CL/NLP efforts often start and end with the orthographic form of words (exceptions include speech recognition and speech synthesis). For Arabic, this means dealing with noise and ambiguity not just in Arabic script, but potentially in other scripts such as the Arabizi romanizations common in social media (Al-Badrashiny et al. 2014). To address Arabic orthographic noise, many CL/NLP researchers often normalize the various forms of Hamzated Alif, mapping أ and إ to ا and converting the Alif-Maqsura ى to ي. The Qatar Arabic Language Bank (QALB) project created a 2-million-word resource for training MSA spelling-correction systems, and set up a shared task around it (Zaghouani et al. 2014; Mohit et al. 2014). For dialectal Arabic, CODA (Habash et al. 2012b) has gained a lot of support since its inception: it has been used as part of annotation projects for Egyptian Arabic (Maamouri et al. 2014) and Palestinian Arabic (Jarrar et al. 2017); it was also used as part of efforts on spelling correction (Eskander et al. 2013) and automatic transliteration from Arabizi to Arabic script (Al-Badrashiny et al. 2014).[1]

We should note that many Arabic CL/NLP resources and papers use one-to-one orthographic transliterations such as the Buckwalter transliteration (a romanization) when discussing Arabic script orthographic phenomena, such as the various forms of Hamza (Habash et al. 2007). These mappings were originally developed when support for Arabic script was lacking, but they continue to be used today because they are easier to debug, e.g., to detect visually if some diacritic is missing or doubled. These transliteration schemes are only used system-internally. Since there are multiple variants, users are cautioned not to mix them up (Habash 2010).

18.4.3 Morphological Analysis and Disambiguation

Morphological analysis and disambiguation is an area that has received a lot of attention in Arabic CL/NLP. This is understandable given the complexity of Arabic morphology and its interaction with orthographic noise.

[1] Most recently, the Palestinian CODA conventions have been adopted by a website for teaching Colloquial Arabic: www.learnpalestinianarabic.com.

https://doi.org/10.1017/9781108277327.019 Published online by Cambridge University Press

Once Arabic text has been analysed and disambiguated, its complexity becomes comparable to other other languages for syntactic and semantic analysis. Morphological analysis refers specifically to models that identify for a given word all of its analyses out of context. Disambiguation refers to identifying a specific analysis in context. There are many types of analysis of Arabic words with varying degree of depth. Ideally, an analysis should specify the lemma (lexical or dictionary entry), the diacritization, part-of-speech and all the linguistic features. Morphological processing typically also includes tokenization into a number of possible schemes (Habash and Sadat 2006). Morphological ML features in Arabic include a large space that has been explored by many researchers in different applications (Habash and Sadat 2006; Marton et al. 2013; Guzmán et al. 2016).

For MSA, the most commonly used morphological analyser is SAMA (Graff et al. 2009), which is used as part of the Penn Arabic Treebank (Maamouri et al. 2004), and the MADAMIRA system for analysis and disambiguation Pasha et al. (2014). The MADAMIRA results have been recently surpassed by extending its approach with neural models (Zalmout and Habash 2017). Other Arabic morphology systems include Smrž (2007), Boudchiche et al. (2017), and Abdelali et al. (2016). For dialectal Arabic, there is relatively less work on morphology. Habash et al. (2012a) and Khalifa et al. (2017) have developed analysers for Egyptian Arabic and Gulf Arabic, respectively. Eskander et al. (2016) described and implemented a paradigm-completion approach to build dialectal morphological analysers from annotated data.

18.4.4 Syntactic Parsing

Syntactic parsing is an enabling technology that assigns a syntactic structure to a sequence of words, e.g., identifying the span and relationship of a noun phrase to a verb. Syntactic parsing is used for a variety of higher-order CL/NLP applications such as machine translation and automatic summarization.

The dominant approaches to syntactic parsing rely on machine learning of parsing models from treebanks, or collections of sentences paired with their syntactic analyses. For Arabic, there are a number of such treebanks: the Penn Arabic Treebank (PATB), which uses a phrase-structure representation (Maamouri et al. 2004), the Prague Arabic Dependency Treebank (PADT) (Smrž et al. 2008), and the Columbia Arabic Treebank (CATiB) (Habash and Roth 2009), both of which use different dependency representations. The Quran Corpus Treebank (Dukes and Buckwalter 2010) uses a hybrid representation inspired by traditional Arabic grammar. The PATB representation has been mapped into a lexical functional grammar (LFG) representation (Tounsi et al. 2009) as well as into the universal dependency (UD) project representation (Nivre et al. 2016; Taji et al. 2017). Other than the Quran Corpus, MSA treebanks focus on the news genre, although there are some recent efforts in treebanking Arabic dialects (Maamouri et al. 2014).

https://doi.org/10.1017/9781108277327.019 Published online by Cambridge University Press

There are several state-of-the-art Arabic parsers, including the Stanford Arabic Parser (PATB phrase structure) (Green and Manning 2010) and the CAMeL Parser (CATiB dependency) (Shahrour et al. 2016). A shallower version of syntactic parsing is base-phrase chunking (BPC), where sequences of adjacent words are grouped together to form syntactic phrases such as NPs and VPs (Diab 2007). The Arabic analysis tool MADAMIRA (Pasha et al. 2014) includes BPC among its outputs. Chiang et al. (2006) explored a number of techniques to optimize the use of MSA treebanks for Arabic dialect parsing.

18.4.5 Semantic Modelling

We describe three efforts that fall under semantic modelling.

First is the development of the Arabic WordNet (Elkateb et al. 2006). WordNets are machine-readable lexical databases that group words into clusters of synonyms called synsets, each of which is thought of as representing a unique word sense. After the first WordNet (in English) (Fellbaum 1998), many wordNets were created and linked to the English WordNet. This is the same for Arabic. Although the size of Arabic WordNet is much smaller than the English WordNet, it has been used to bootstrap resources by exploiting existing annotations linked to the English WordNet (Badaro et al. 2014).

Second is the development of the Arabic proposition bank (propbank). A propbank is a semantically annotated corpus, where propositions and their arguments are marked in the form of predicate-argument information and semantic-role labels on top of an existing syntactic treebank (Zaghouani et al. 2010).

Finally, there has been a lot of work on sentiment analysis in Arabic, where the target is typically to assign a sentence a label indicating whether it is positive, negative, or neutral. This work includes the development of various resources and lexicons mapping words or lemmas to their sentiment (Abdul-Mageed and Diab 2012; Badaro et al. 2014), as well as developing systems for sentiment detection (Abdul-Mageed et al. 2012; Al Sallab et al. 2015).

18.5 Initiation Resources for Arabic Computational Linguistics

In this section, we present a number of resources that are good starting points for linguists interested in expanding into CL/NLP, in general, and Arabic CL/NLP specifically.

18.5.1 Where to Start

CL/NLP is an interdisciplinary field where collaborations among linguists and computer scientists are common and very productive. Researchers from both

https://doi.org/10.1017/9781108277327.019 Published online by Cambridge University Press

sides coming together benefit greatly from learning as much as possible about each other's fields. For linguists, learning to program is extremely helpful, and a necessity for pursuing deeper research efforts in CL/NLP . There are many online resources for learning how to program. For a *gentle introduction for linguists*, see (Hovy 2012). That said, the best route for a linguist interested in CL/NLP is to take courses, preferably as part of a graduate degree, in one of the many Computational Linguistics graduate programmes. There are also many online courses, e.g., Coursera's Natural Language Processing course.[2]

18.5.2 Resources

Books Popular CL/NLP textbooks include Manning and Schutze (1999) and Jurafsky and Martin (2008). The Synthesis Lectures on Human Language Technologies (Hirst 2008–2017) include a number of publications on CL/NLP focusing on specific topics. In that series, there is a book on Arabic natural language processing (Habash 2010). A number of tutorials on Arabic NLP are available online.[3]

Tools and Data The top repositories for CL/NLP research resources, specifically data, are the Linguistic Data Consortium (LDC)[4] and the European Language Resources Association (ELRA).[5] The LDC has over 190 Arabic resources in its catalogue, including the highly used Penn Arabic Treebank (Maamouri et al. 2004) and Standard Arabic Morphological Analyzer (Graff et al. 2009). ELRA has over a hundred Arabic resources in its catalogue. For lists of commonly used corpora, lexicons, and tools, including freely available resources, see Habash (2010); Zaghouani (2014); Shoufan and Alameri (2015).

Publication Venues and Archives Conferences are the primary venues of publication in CL/NLP, where the top peer-reviewed conferences are very competitive (≈25 per cent acceptance rate). Conferences typically expect four- to eight-page anonymous paper submissions. The proceedings of the conferences of the Association for Computational Linguistics (ACL) and its various chapters[6] are publicly archived as part of the ACL anthology,[7] which also hosts some non-ACL conferences. Most notable among the non-ACL conferences are COLING (International Conference on Computational Linguistics) and LREC (Language Resources and Evaluation Conference).[8] Workshops (as part of conferences) and shared task competitions are common, and typically publish their proceedings. There are a number of journals in the field, e.g., *Computational Linguistics*,[9] *Transactions of the ACL*,[10], *Computer Speech and Language*,[11] and *Language*

[2] https://www.coursera.org/learn/language-processing. [3] www.nizarhabash.com/teaching.
[4] www.ldc.upenn.edu/. [5] http://catalog.elra.info/. [6] www.aclweb.org/. [7] http://aclweb.org/anthology/.
[8] www.lrec-conf.org/. [9] www.mitpressjournals.org/loi/coli. [10] www.transacl.org/ojs/index.php/tacl.
[11] www.journals.elsevier.com/computer-speech-and-language.

https://doi.org/10.1017/9781108277327.019 Published online by Cambridge University Press

Resources and Evaluation.[12] As with other areas in the sciences, arXiv.org has become the place to pre-publish the latest works in CL/NLP.[13]

Arabic CL/NLP papers are often published in the international venues mentioned above; however, a number of Arabic CL/NLP workshops and conferences took place in the early twenty-first century, both in the Arab world and in association with international conferences. Examples include the Arabic Natural Language Processing Workshop, typically co-located with an ACL event (2014, 2015, and 2017), the Workshop on Arabic Corpora and Processing Tools, typically co-located with LREC (2014, 2016, 2018), and the International Conference on Arabic Computational Linguistics (2015, 2016, 2017).

18.6 The Future of Arabic Computational Linguistics

The field of CL/NLP as a whole saw some great growth in the first two decades of the twenty-first century (Table 18.1). This progress is a result of a growing competitive community that has a strong collaborative culture around sharing resources (data and tools) and aiming for language-independent solutions. Today the results of this progress can be seen in terms of robust speech recognition technology, machine translation, and dialogue systems making it into products such as Siri and Google Translate. With this market growth comes increased customer expectations, which fuel competition and encourage innovation. This suggests more jobs in CL/NLP will be created, not only to develop specific applications, but also to create data annotations to train and evaluate said applications.

The field of Arabic CL/NLP naturally benefits from general progress in CL/NLP since most of the developed techniques are language independent. Furthermore, the commercial growth around English-language technologies only helps raise demand for support for other languages, as large companies are continuously trying to get access into more regions around the world. Arabic is currently not as present among the various languages with ready-to-use technologies, but this is coming and it is not an issue of *if* but rather *when*.

Research and work on Arabic CL/NLP specifically has always lagged behind English and other languages. If we consider the data in Table 18.1,[14] we see

[12] https://link.springer.com/journal/10579. [13] https://arxiv.org/list/cs.CL/recent.

[14] The results in Table 18.1 are based on Google Scholar counts for the query 'natural language processing' and the name of a language (unquoted) such as *English* or *Arabic dialect*. The counts are collected for each year separately and then aggregated. The search was conducted on 1 December 2017, and did not include patents or citations. Next to all languages except English, a ratio of the number of publications compared to that of English ($_{en}$) or Arabic ($_{ar}$) is specified. Since these numbers are counts of Google Scholar search results, they are rough estimates, capturing only when these languages are mentioned, even if no work was actually done on them as part of the publication. Finally, we acknowledge the self-imposed limitation of only considering documents written in English. We justified this choice by the fact that English is the primary publishing language in the field of CL/NLP.

https://doi.org/10.1017/9781108277327.019 Published online by Cambridge University Press

Table 18.1 *A summary of the number of publications in 'natural language processing' for a number of languages between 1997 and 2016*

Year	English	Chinese		German		Hindi		Arabic		Arabic Dialects	
1997–2001	8,380	2,413	$(29\%)_{en}$	4,261	$(51\%)_{en}$	142	$(2\%)_{en}$	523	$(6\%)_{en}$	101	$(19\%)_{ar}$
2002–2006	18,530	7,850	$(42\%)_{en}$	9,770	$(53\%)_{en}$	554	$(3\%)_{en}$	1,842	$(10\%)_{en}$	336	$(18\%)_{ar}$
2007–2011	34,110	19,690	$(58\%)_{en}$	17,470	$(51\%)_{en}$	1,889	$(6\%)_{en}$	5,118	$(15\%)_{en}$	822	$(16\%)_{ar}$
2012–2016	58,540	35,650	$(61\%)_{en}$	27,900	$(48\%)_{en}$	3,992	$(7\%)_{en}$	9,250	$(16\%)_{en}$	1,801	$(19\%)_{ar}$

some patterns that can give some insights into where Arabic CL/NLP is and where it is heading.

First, Arabic and Arabic dialect CL/NLP have grown a lot, with the total number of publications in 2012–16 being ~18 times the number in 1997–2001, compared to English and German (~7 times).

Second, Arabic CL/NLP grew particularly fast between 2002 and 2012. When considering relative publication growth as the ratio to English publication (to offer a normalizing reference), we see that Arabic CL/NLP publications were about 6 per cent of English CL/NLP publications in 1997–2001. However this ratio rose to 10 per cent for 2002–2006 and again to 15 per cent in 2007–2011. But this relative growth slowed down in the period 2012–2016. The historical rise in work on and interest in Arabic CL/NLP is probably due to major funding for Arabic-related research, specifically machine translation in the USA after the events of 11 September 2001. This includes almost a decade of sustained funding under the DARPA programmes (2002~2013): GALE (Global Autonomous Language Exploitation), BOLT (Broad Operational Language Translation), and others. These projects involved a large number of researchers from industry and academia, and created a large number of resources, many of which were made available for research purposes through the LDC. With the end of such large projects, smaller projects such as those funded by the Qatar National Research Fund became more the norm, e.g., QALB (Qatar Arabic Language Bank)[15] and OMA (Opinion Mining for Arabic).[16] Also, after that period came the creation of institutionally funded Arabic-focused publication-targeting research labs such as the Qatar Computing Research Institute's Arabic Language Technology group and New York University Abu Dhabi's Computational Approaches to Modeling Language (CAMeL) lab. For comparison purposes, German CL/NLP, e.g., remained at about 50 per cent of the size of English CL/NLP from 1997 to 2016. Chinese CL/NLP by comparison leaped the fastest from 24 per cent to 62 per cent, while Hindi CL/NLP is the slowest among the languages we compared.

Third, Arabic dialects have become a more prominent part of Arabic CL/NLP. The percentage of Arabic dialect publications in 1997 was only 11 per cent of all Arabic CL/NLP; but it had risen to 22 per cent in 2016. This is partly due to a sense that work on certain aspects of MSA has little space for progress, particularly in the area of morphology. Additionally, a big motivation for dialect CL/NLP is the growing interest in language technologies for opinion mining and translation in social media, which features dialectal Arabic more than MSA.

Overall, Arabic CL/NLP has made big leaps in relative growth. But it can be said to be fifteen years behind English, since in 2016 the Arabic CL/NLP publication count was comparable to that for English in 2001. And today,

[15] http://nlp.qatar.cmu.edu/qalb/. [16] http://oma-project.azurewebsites.net/.

https://doi.org/10.1017/9781108277327.019 Published online by Cambridge University Press

Arabic CL/NLP seems to be plateauing at 15 per cent relative to English. For Arabic CL/NLP to push forward (following Chinese CL/NLP's example), more has to be done: more training of CL/NLP people (computer scientists and linguists), and more tools and data, not just created, but also made publicly available to other researchers. All of these issues depend on funding for research and development. The possible sources of funding may have to be governmental programmes and, eventually, industry-led initiatives. As things stand today, the market for Arabic CL/NLP in the Arab world is nowhere near its full potential, but its growth is inevitable. It will be very interesting to see how Arabic CL/NLP will shape up over the next decade. We predict the growth in commercialization for English CL/NLP technologies will be contagious and spread to the Arab world; and it will most likely change how Arabs think of their language and its relationship to technology. For example, we predict that Arabic speakers will develop higher expectations of being able to use their dialect ('normal speech') to directly communicate with their machines. We also predict that the attitude of expecting the use of MSA in personal technological contexts will fall by the wayside as different companies will compete for wider market share by accommodating users' needs and capturing their interest.

18.7 Summary

In summary, we have introduced Arabic computational linguistics and situated it within the field of Computational Linguistics / Natural Language Processing. Arabic brings to CL/NLP a combination of many interesting and tough challenges. Although a lot of progress has taken place in the field, there is still much more work to do. The future of the field promises a lot of possibilities for, and opportunities in, language technology development.

References

Abdelali, A., Darwish, K., Durrani, N., and Mubarak, H. (2016). Farasa: A fast and furious segmenter for Arabic. In *Proceedings of the Meeting of the North America Association for Computational Linguistics (NAACL)*. San Diego, California.

Abdul-Mageed, M. and Diab, M. (2012). Toward building a large-scale Arabic sentiment lexicon. In *Proceedings of The International Global WordNet Conference*. Matsue, Japan.

Abdul-Mageed, M., Kuebler, S., and Diab, M. (2012). SAMAR: A system for subjectivity and sentiment analysis of Arabic social media. In *Proceedings of the Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*. Jeju, Korea.

https://doi.org/10.1017/9781108277327.019 Published online by Cambridge University Press

Al-Badrashiny, M., Eskander, R., Habash, N., and Rambow, O. (2014). Automatic transliteration of Romanized dialectal Arabic. In *Proceedings of the Conference on Computational Natural Language Learning*. Ann Arbor, Michigan.

Al Sallab, A. A., Baly, R., Badaro, G., Hajj, H., El Hajj, W., and Shaban, K. B. (2015). Deep learning models for sentiment analysis in Arabic. In *Proceedings of the Arabic Natural Language Processing Workshop (WANLP)*. Beijing, China.

Badaro, G., Baly, R., Hajj, H., Habash, N., and El Hajj, W. (2014). A large scale Arabic sentiment lexicon for Arabic opinion mining. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*. Doha, Qatar, 165–73.

Bouamor, H., Habash, N., and Oflazer, K. (2014). A multidialectal parallel corpus of Arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. Reykjavik, Iceland.

Boudchiche, M., Mazroui, A., Bebah, M. O. A. O., Lakhouaja, A., and Boudlal, A. (2017). AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer. *Journal of King Saud University – Computer and Information Sciences*, 29(2), 141–6.

Chiang, D., Diab, M., Habash, N., Rambow, O., and Shareef, S. (2006). Parsing Arabic dialects. In *Proceedings of the Meeting of the European Association for Computational Linguistics (EACL)*. Trento, Italy.

Diab, M. (2007). Improved Arabic base phrase chunking with a new enriched POS tag set. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages (CASL)*. Prague, Czech Republic.

Dukes, K., and Buckwalter, T. (2010). A dependency treebank of the Quran using traditional Arabic grammar. In *Proceedings of the International Conference on Informatics and Systems (INFOS)*. Cairo, Egypt.

Dukes, K., Atwell, E., and Habash, N. (2013). Supervised collaboration for syntactic annotation of Quranic Arabic. In *Language Resources and Evaluation*, 47(1), 33–62.

El Kholy, A. and Habash, N. (2012). Orthographic and morphological processing for English–Arabic statistical machine translation. *Machine Translation*, 26(1–2), 25–45.

El Kholy, A. and Habash, N. (2015). Morphological constraints for phrase pivot statistical machine translation. In *Proceedings of the Machine Translation Summit (MTSummit)*. Miami, Florida.

Elfardy, H. and Diab, M. (2013). Sentence-level dialect identification in Arabic. In *Proceedings of the Association for Computational Linguistics*. Sofia, Bulgaria.

Elkateb, S., Black, W., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., et al. (2006). Building a WordNet for Arabic. In *Proceedings of the International Conference on Language Resources and Evaluation*. Genoa, Italy.

Eskander, R., Habash, N., Rambow, O., and Tomeh, N. (2013). Processing spontaneous orthography. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*. Atlanta, Georgia.

https://doi.org/10.1017/9781108277327.019 Published online by Cambridge University Press

Eskander, R., Habash, N., Rambow, O., and Pasha, A. (2016). Creating resources for dialectal Arabic from a single annotation: A case study on Egyptian and Levantine. In *Proceedings of the International Conference on Computational Linguistic (COLING)*. Osaka, Japan.

Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Graff, D., Maamouri, M., Bouziri, B., Krouna, S., Kulick, S., and Buckwalter, T. (2009). *Standard Arabic Morphological Analyzer – Version 3.1 Catalog No.: LDC2009E73*. Linguistic Data Consortium, University of Pennsylvania.

Green, S. and Manning, C. D. (2010). Better Arabic parsing: Baselines, evaluations, and analysis. *In Proceedings of the 23rd International Conference on Computational Linguistics*. Beijing, China, 394–402.

Guzmán, F., Bouamor, H., Baly, R., and Habash, N. (2016). Machine translation evaluation for Arabic using morphologically-enriched embeddings. In *Proceedings of COLING 2106*. Osaka, Japan.

Habash, N. (2010). *Introduction to Arabic Natural Language Processing*, vol. 3. Morgan & Claypool.

Habash, N. and Roth, R. (2009). CATiB: The Columbia Arabic Treebank. In *Proceedings of the ACL-JNLP Conference*. Suntec, Singapore, 221–4.

Habash, N. and Sadat, F. (2006). Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*. New York.

Habash, N., Soudi, A., and Buckwalter, T. (2007). On Arabic transliteration. In A. Soudi, G. Neumann, and A. van den Bosch, eds., *Arabic Computational Morphology: Text, Speech and Language Technology*, vol. 38. Dordrecht: Springer, 15–22.

Habash, N., Eskander, R., and Hawwari, A. (2012a). A morphological analyzer for Egyptian Arabic. In *Proceedings of the Workshop on Computational Morphology and Phonology*. Montréal, Canada.

Habash, N., Diab, M., and Rambow, O. (2012b). Conventional orthography for dialectal Arabic. In *Proceedings of the International Conference on Language Resources and Evaluation*. Istanbul, Turkey.

Habash, N., Zalmout, N., Taji, D., Hoang, H., and Alzate, M. (2017). A parallel corpus for evaluating machine translation between Arabic and European languages. In *Proceedings of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain.

Hirst, G. (ed.) (2008–2017). *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool.

Hovy, D. (2012). *Programming in Python for Linguists: A Gentle Introduction*. www.dirkhovy.com/portfolio/papers/download/pfl_handout.pdf; last accessed 10 December 2020.

Jarrar, M., Habash, N., Alrimawi, F., Akra, D., and Zalmout, N. (2017). Curras: An annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, 51, 745–75.

https://doi.org/10.1017/9781108277327.019 Published online by Cambridge University Press

Jinxi, X. (2002). *UN Parallel Text (Arabic-English), LDC Catalog No.: LDC2002E15*. Linguistic Data Consortium, University of Pennsylvania.

Jurafsky, D. and Martin, J. H. (2008). *Speech and Language Processing*. Upper Saddle River, NJ: Prentice Hall.

Khalifa, S., Habash, N., Abdulrahim, D., and Hassan, S. (2016). A large scale corpus of Gulf Arabic. In *Proceedings of the Language Resources and Evaluation Conference 2016*. Portorož, Slovenia.

Khalifa, S., Hassan, S., and Habash, N. (2017). A morphological analyzer for Gulf Arabic verbs. In *Proceedings of the Third Arabic Natural Language Processing Workshop (WANLP)*. Valencia, Spain, 35–45.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*. Phuket, Thailand. 79–86.

Maamouri, M., Bies, A., Buckwalter, T., and Mekki, W. (2004). The Penn Arabic Treebank: Building a large-scale annotated Arabic corpus. In *Proceedings of the NEMLAR Conference on Arabic Language Resources and Tools*. Cairo, Egypt.

Maamouri, M., Bies, A., Kulick, S., Ciul, M., Habash, N., and Eskander, R. (2014). Developing an Egyptian Arabic Treebank: Impact of dialectal morphology on annotation and tool development. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. Reykjavik, Iceland.

Manning, C. and Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.

Marton, Y., Habash, N., and Rambow, O. (2013). Dependency parsing of Modern Standard Arabic with lexical and inflectional features. *Computational Linguistics*, 39(1), 161–94.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mohit, B., Rozovskaya, A., Habash, N., Zaghouani, W., and Obeid, O. (2014). The first QALB shared task on automatic text correction for Arabic. In *Proceedings of the Arabic Natural Language Processing Workshop (WANLP)*. Doha, Qatar.

Munteanu, D. S. and Marcu, D. (2007). *ISI Arabic–English Automatically Extracted Parallel Tex. Catalog No.: LDC2007T08*. Linguistic Data Consortium, University of Pennsylvania.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., et al. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of International Conference on Language Resources and Evaluation*. Portorož, Slovenia.

Pasha, A., Al-Badrashiny, M., El Kholy, A., Eskander, R., Diab, M., Habash, N., et al. (2014). MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the International Conference on Language Resources and Evaluation*. Reykjavik, Iceland.

Rafalovitch, A. and Dale, R. (2009). United Nations General Assembly Resolutions: A six-language parallel corpus. In *Proceedings of the 12th Machine Translation Summit*. Ottawa, Canada.

Salloum, W. and Habash, N. (2011). Dialectal to Standard Arabic paraphrasing to improve Arabic–English statistical machine translation. In *Proceedings of the Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*. Edinburgh, UK.

Shahrour, A., Khalifa, S., Taji, D., and Habash, N. (2016). CamelParser: A system for Arabic syntactic analysis and morphological disambiguation. *In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*. Osaka, Japan, 228–32.

Shoufan, A. and Alameri, S. (2015). Natural language processing for dialectical Arabic: A survey. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*. Beijing, China, 36–48.

Smrž, O. (2007). ElixirFM: Implementation of functional Arabic morphology. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*. Prague, Czech Republic, 1–8.

Smrž, O., Bielický, V., Kouřilová, I., Kráčmar, J., Hajič, J., and Zemánek, P. (2008). Prague Arabic Dependency Treebank: A word on the million words. In *Proceedings of the International Conference on Language Resources and Evaluation*. Marrakech, Morocco.

Taji, D., Habash, N., and Zeman, D. (2017). Universal dependencies for Arabic. In *Proceedings of the Third Arabic Natural Language Processing Workshop*. Valencia, Spain, 166–76.

Tounsi, L., Attia, M., and van Genabith, J. (2009). Automatic treebank-based acquisition of Arabic LFG dependency structures. In *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*. Athens, Greece, 45–52.

Watson, J. C. E. (2007). *The Phonology and Morphology of Arabic*. Oxford: Oxford University Press.

Zaghouani, W. (2014). Critical survey of the freely available Arabic corpora. In *Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools*. Reykjavik, Iceland.

Zaghouani, W., Diab, M., Mansouri, A., Pradhan, S., and Palmer, M. (2010). The Revised Arabic Propbank. In *Proceedings of the Linguistic Annotation Workshop*. Uppsala, Sweden.

Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh, N., Rozovskaya, A., et al. (2014). Large-scale Arabic error annotation: Guidelines and framework. In *Proceedings of the International Conference on Language Resources and Evaluation* . Reykjavik, Iceland.

Zalmout, N. and Habash, N. (2017). Don't throw those morphological analyzers away just yet: Neural morphological disambiguation for Arabic. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark.

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., et al. (2012). Machine translation of Arabic dialects. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*. Montréal, Canada.

https://doi.org/10.1017/9781108277327.019 Published online by Cambridge University Press