Preface

Modern communication technologies, such as the television and the Internet, have made readily available massive amounts of information in many languages. More such data is being generated in real time, 24 h a day and 7 days a week, aided by social networking sites such as Facebook and Twitter. This information explosion is in the form of multilingual audio, video, and Web content. The task of processing this large amount of information demands effective, scalable, multilingual media processing, monitoring, indexing, and search solutions. Natural Language Processing (NLP) technologies have long been used to address this task, and several researchers have developed several technical solutions for it. In the last two decades, NLP researchers have developed exciting algorithms for processing large amounts of text in many different languages. Nowadays the English language has obtained the lion's share in terms of available resources as well as developed NLP technical solutions. In this book, we address another group of interesting and challenging languages for NLP research, that is, the Semitic languages. The Semitic languages have existed in written form since a very early date, with texts written in a script adapted from Sumerian cuneiform. Most scripts used to write Semitic languages are *abjads*, a type of alphabetic script that omits some or all of the vowels. This is feasible for these languages because the consonants in the Semitic languages are the primary carriers of meaning. Semitic languages have interesting morphology, where word roots are not themselves syllables or words, but isolated sets of consonants (usually three characters). Words are composed out of roots by adding vowels to the root consonants (although prefixes and suffixes are often added as well). For example, in Arabic, the root meaning "write" has the form k - t - b. From this root, words are formed by filling in the vowels, e.g., kitAb "book," kutub "books," kAtib "writer," kuttAb "writers," kataba "he wrote," yaktubu "he writes," etc. Semitic languages, as stated in Wikipedia, are spoken by more than 270 million people. The most widely spoken Semitic languages today are Arabic (206 million native speakers), Amharic (27 million), Hebrew (7 million), Tigrinya (6.7 million), Syriac (1 million), and Maltese (419 thousand). NLP research applied to Semitic languages has been the focus of attention of many researchers for more than a decade, and several technical solutions have been proposed, especially

Arabic NLP where we find a very large amount of accomplished research. This will be reflected in this book, where Arabic will take the lion's share. Hebrew also has been the center of attention of several NLP research works, but to a smaller degree when compared to Arabic. Most of the key published research works in Hebrew NLP will be discussed in this book. For Amharic, Maltese, and Syriac, because of the very limited amount of NLP research publicly available, we didn't limit ourselves to present key techniques, but we also proposed solutions inspired from Arabic and Hebrew. Our aim for this book is to provide a "one-stop shop" to all the requisite background and practical advice when building NLP applications for Semitic languages. While this is quite a tall order, we hope that, at a minimum, you find this book a useful resource.

Similar to English, the dominant approach in NLP for Semitic languages has been to build a statistical model that can learn from examples. In this way, a model can be robust to changes in the type of text and even the language of text on which it operates. With the right design choices, the same model can be trained to work in a new domain simply by providing new examples in that domain. This approach also obviates the need for researchers to lay out, in a painstaking fashion, all the rules that govern the problem at hand and the manner in which those rules must be combined. A statistical system typically allows for researchers to provide an abstract expression of possible *features* of the input, where the relative importance of those features can be learned during the *training* phase and can be applied to new text during the *decoding*, or *inference*, phase. While this book will devote some attention to cutting-edge algorithms and techniques, the primary purpose will be a thorough explication of best practices in the field. Furthermore, every chapter describes how the techniques discussed apply to Semitic languages.

This book is divided into two parts. Part I, includes the first five chapters and lays out several core NLP problems and algorithms to attack those problems. The first chapter introduces some basic linguistic facts about Semitic languages, covering orthography, morphology, and syntax. It also shows a contrastive analysis of some of these linguistic phenomena across the various languages. The second chapter introduces the important concept of *morphology*, the study of the structure of words, and ways to process the diverse array of morphologies present in Semitic languages. Chapter 3 investigates the various methods of uncovering a sentence's internal structure, or *syntax*. Syntax has long been a dominant area of research in NLP. This dominance is explained in part by the fact that the structure of a sentence is related to the sentence's meaning, and so uncovering syntactic structure can serve as a first step toward "understanding" a sentence. One step beyond syntactic parsing toward understanding a sentence is to perform semantic parsing that consists in finding a structured meaning representation for a sentence or a snippet of text. This is the focus of Chap. 4 that also covers a related subproblem known as *semantic role labeling*, which attempts to find the syntactic phrases that constitute the *arguments* to some verb or *predicate*. By identifying and classifying a verb's arguments, we come closer to producing a *logical form* for a sentence, which is one way to represent a sentence's meaning in such a way as to be readily processed by machine. In several NLP applications, one simply wants to accurately estimate the likelihood

of a word (or word sequence) in a phrase or sentence, without the need to analyze syntactic or semantic structure. The history, or context, that is used to make that estimation might be long or short, knowledge rich, or knowledge poor. The problem of producing a likelihood or probability estimate for a word is known as *language modeling*, and is the subject of Chap. 5.

Part II, takes the various core areas of NLP described in Part I and explains how they are applied to real-world NLP applications available nowadays for several Semitic languages. Chapters in this part of the book explore several tradeoffs in making various algorithmic and design choices when building a robust NLP application for Semitic languages, mainly Arabic and Hebrew. Chapter 6 describes one of the oldest problems in the field, namely, *machine translation*. Automatically translating from one language to another has long been a holy grail of NLP research, and in recent years the community has developed techniques that make machine translation a practical reality, reaping rewards after decades of effort. This chapter discusses recent efforts and techniques in translating Semitic languages such as Arabic and Hebrew to and from English.

The following three chapters focus on the core parts of a larger application area known as *information extraction*. Chapter 7, describes ways to identify and classify *named entities* in text. Chapter 8, discusses the linguistic relation between two textual entities which is determined when a textual entity (the anaphor) refers to another entity of the text which usually occurs before it (the antecedent). The last chapter of this trilogy, Chap. 9, continues the information extraction discussion, exploring techniques for finding out how two entities are related to each other, known as *relation extraction*.

The subject of finding few documents or subparts of documents that are relevant based on a search query is clearly an important NLP problem, as it shows the popularity of search engines such as Bing or Google. This problem is known as *information retrieval* and is the subject of Chap. 10. Another way in which we might tackle the sheer quantity of text is by automatically summarizing it. This is the content of Chap. 11. This problem either involves finding the sentences, or bits of sentences, that contribute toward providing a relevant summary, or else ingesting the text, summarizing its meaning in some internal representation, and then *generating* the text that constitutes a summary. Often, humans would like machines to process text automatically because they have questions they seek to answer. These questions can range from simple, factoid-like questions, such as "what is the family of languages to which Hebrew belongs?" to more complex questions such as "what political events succeeded the Tunisian revolution?" Chapter 12 discusses recent techniques to build systems to answer these types of questions automatically.

In several cases, we would like our speech to be automatically transcribed so that we can interact more naturally with machines. The process of converting speech into text, known as *Automatic Speech Recognition*, is the subject of Chap. 13. Plenty of advances have been made in the recent years in Arabic speech recognition. Current systems for Arabic and Hebrew achieve very reasonable performance and can be used in real NLP applications.

As much as we hope this book is self-contained and covers most research work in NLP for Semitic languages, we also hope that for you, the reader, it serves as the beginning and not an end. Each chapter has a long list of relevant work upon which it is based, allowing you to explore any subtopic in great detail. The large community of NLP researchers is growing throughout the world, and we hope you join us in our exciting efforts to process language automatically and that you interact with us at universities, at research labs, at conferences, on social networks, and elsewhere. NLP systems of the future and their application to Semitic languages are going to be even more exciting than the ones we have now, and we look forward to all your contributions!

Bellevue WA, USA   Imed Zitouni
March 2014

Springer