Chapter 22
Morphological Analysis of Ill-Formed Arabic Verbs for Second Language Learners

Khaled Shaalan
The British University in Dubai, UAE

Marwa Magdy
Cairo University, Egypt

Aly Fahmy
Cairo University, Egypt

ABSTRACT
Arabic is a language of rich and complex morphology. The nature and peculiarity of Arabic make its morphological and phonological rules confusing for second language learners (SLLs). The conjugation of Arabic verbs is central to the formulation of an Arabic sentence because of its richness of form and meaning. In this research, we address issues related to the morphological analysis of ill-formed Arabic verbs in order to identify the source of errors and provide an informative feedback to SLLs of Arabic. The edit distance and constraint relaxation techniques are used to demonstrate the capability of the proposed system in generating all possible analyses of erroneous Arabic verbs written by SLLs. Filtering mechanisms are applied to exclude the irrelevant constructions and determine the target stem which is used as the base for constructing the feedback to the learner. The proposed system has been developed and effectively evaluated using real test data. It achieved satisfactory results in terms of the recall rate.

INTRODUCTION
Language is a way of communicating ideas and feelings among people by the use of conventional symbols. People need to learn second languages to be able to communicate with other non-native speakers. Second language acquisition is a difficult task, especially for adults. There are various methods to acquire a new language and all of them require some form of feedback, which can be described as a reaction to what has been said or written. This feedback most often comes from other human beings with whom the language
learner is interacting. There are, however, other means to receive automated feedback. One is the use of intelligent language tutoring system (ILTS) software. This software contains exercises for language learners. Their response to these exercises is analyzed by the system which provides some form of feedback that could identify the exact source of error a learner has made.

There are some types of exercises that are easy to be error diagnosed, such as multiple choice questions and gap filling exercises, because the number of possible answers is very limited. Simple methods can then be employed to provide a feedback to learners. Whenever the range of possible answers is large or even infinite, specialized intelligent tools are needed. For instance, in the case of exercises requiring learners to produce sentences in the language they are learning, Natural Language Processing (NLP) tools and techniques are necessary to analyze the learner’s answer and produce intelligent feedback. In a morphological rich language such Arabic, an inflected verb can form a complete sentence (e.g. the verb /samiEtuka/ [heard-I-you]) contains a complete syntactic structure in just a one-word sentence. In this case the NLP tools and techniques are also required to analyze the learner’s answer and produce intelligent feedback.

The work presented in this chapter addresses issues related to the morphological analysis of ill-formed Arabic verbs written by beginner to intermediate SLLs. The proposed system is an integral part of an ILTS for Arabic. SLLs of Arabic, however, face a lot of morphological and syntactic difficulties during their language learning tasks, such as word formation, word recognition, sentence construction, and disambiguation. This complexity in learning Arabic makes addressing the diagnosis of Arabic lexical errors a challenge. This has motivated us to develop a tool that addresses the word formation problem that is usually faced by SLLs of Arabic. This is achieved by making the proposed tool analyzes the learner’s answer which is used to provide learner with some form of feedback that identifies the exact source of the error s/he might made.

The edit distance and constraint relaxation techniques are used to generate all possible analyses of erroneous Arabic verbs. Filtering mechanisms are applied after the extraction of affixes and stems to exclude the irrelevant constructions and determine the target stem. For each case, a morphological gloss is incrementally formulated which is to be used as a base for constructing the feedback to the learner.

Many research, however, have attacked the problem of Arabic morphological analysis (Ahmed 2000; Beesley 2001; Buckwalter 2002; Darwish 2002; Al-Sughaiyer and Al-Kharashi 2004; Attia 2006). But to the best of our knowledge few research have addressed the problem of analysis of ill-formed Arabic words (e.g., Bowden and Kiraz 1995; Ahmed 2000; Buckwalter 2002). Bowden and Kiraz (1995) investigated the problem of correcting words in Semitic languages including Arabic language. Their approach integrated with morphological analysis using a multi-tape formalism. The model had two-level error rules that handle the following error types: vowel shift, deleted consonant, deleted long vowels, and substituted consonant. Moreover, Ahmed (2000) and Buckwalter (2002) applied some spelling relaxation rules (to deal with orthographic variations like the use of the final letter /h/ instead of the letter /p/) to get all possible analyses of an erroneous word. However, these systems only handle performance errors made by native speakers of Arabic. In contrast to the proposed system that handles competence errors made by nonnative speakers of Arabic. It does so by incorporating morphological knowledge and non-native intuitions into its algorithm. It does not depend on simple string matching between correct and erroneous words.

The rest of this paper is structured as follows. Section 2 discusses lexical error analysis prob-