Review of Doctoral Thesis

Functional Arabic Morphology: Formal System and Implementation *by* Otakar Smrž

Reviewer: Nizar Habash, Ph.D. (Columbia University)

Summary

This thesis presents a formulation of Arabic's complex morphology within a functional approach that anticipates the issues to be addressed in linking morphology and syntax in Arabic. The thesis's contribution is primarily theoretical in nature. However, it is also supplemented with ElixirFM, a complete implementation (not a toy prototype system) that builds on a previous implementation of Arabic morphology, namely the Buckwalter Arabic Morphological Analyzer (BAMA). The implementation takes the shallow, surface-based system of BAMA and extends it in a more meaningful and theoretically coherent manner. In addition, the author provides a suite of related tools and resources that are well motivated for Arabic computational modeling. One is MorphoTrees, a tool for visualizing morphological ambiguity in Arabic words that can be used to speed up annotation. And another is a useful interface with the ArabTEX system for typesetting Arabic.

New Scientific Contributions

For a long time, work on Arabic computational morphology focused on the question of modeling templatic (root and pattern) Semitic morphology. It is now well understood how to do this and how to model shallow surface-to-morph morphology for Arabic. As work on Arabic goes deeper in trying to model syntax for Arabic parsing, the limitation of shallow (illusory) models of Arabic morphology are becoming more apparent. This thesis opens a new page for Arabic computational morphology research by presenting and addressing facts of Arabic's morpho-syntactic interface and using them to guide the design of a functional morphology system. The description of the Arabic facts is correct and complete and the solutions provided to cover them elegant and novel. One example that impressed me is how Arabic definiteness is handled. Here the author presents all the facts including some that have not been addressed completely by previous work. The author's model uses meaningful categories that cover only the possible combinations and naturally eliminates impossible situations. This is in contrast to two other previous works that either are incomplete or have to rely on exclusion rules. The elegant framework presented is very attractive and potentially useful for future issues on Arabic morphology.

The work presented on the visualization tool MorphoTrees is a very interesting way to represent Arabic morphological ambiguity. However, it is not clear how this can be used in an automatic disambiguation system. Of course, its use for human annotation is likely to help speed up the process of annotation and even minimize errors.

One suggestion I have for future publications is including some formal evaluation. There are many interesting motivations that are mentioned in the thesis but that are not revisited or evaluated. For instance, the author could conduct an evaluation on improvements in inter-annotator agreement or annotation speed using MorphoTrees. A simple evaluation of coverage of ElixirFM against the BAMA system is not provided. It is not clear for instance if the re-implementation of BAMA is lossy or not; or if both generation and analysis modes in ElixirFM are completely equivalent or not. A third possibility is to evaluate whether in fact extending the lexicon given this new model is faster than other shallower models. I suspect it is.

Importance to Neighboring Areas

Given Arabic's complex morphology, morphological modeling is a necessary step in any work on Arabic natural language processing. This includes syntactic parsing, machine translation, word sense disambiguation, and natural language generation. This thesis presents a deeper level of representation for Arabic morphology that can be used to provide meaningful features for other tasks. The author addresses the issue of modeling syntax which is very relevant for Arabic parsing. But better morphological representations are necessary for other tasks as well. For instance, Arabic has a very common form of plural called "broken plural". It behaves in many respects like a singular noun since the plurality is created through pattern change and not affixation. Some shallow morphological models are satisfied by treating these plurals as if they are singular nouns. But in the context of functional morphology, we can know that these words are referring to multiple entities. This piece of knowledge is very important in machine translation to languages that explicitly mark plurality such as English or French.

Form of Submitted Thesis

Overall, the thesis is well structured and well written. The author's English is excellent. I only have one suggestion for future presentations of the content: minimize the use of Haskell code in the presentation. The extensive use of Haskell code in the thesis distracts from the points being made. The author in some cases uses hard-to-follow code as the main means of presenting his implementation, when in fact simple tables or pseudo code would have sufficed. For instance, on page 43, the author uses the counter-intuitive three-way Boolean-extended distinction (nothing, just true and just false) to describe the values of the Definite variable instead of using terms comparable to what he uses to denote the values of Case (nominative, accusative or genitive) or Number (singular, dual or plural). In another example, starting in page 90, the author lists three pages of Haskell code to explain spelling rules of Hamza (Arabic letter representing glottal stop) instead of describing in English what rules are implemented and what decisions are made.

Overall Evaluation of Thesis

This thesis is theoretically strong and provides a solid real-world prototype implementation. I believe that the thesis proves the author's ability for creative scientific work.

August 17, 2007

[redacted]

Nizar Habash, Ph.D.
Associate Research Scientist
Center for Computational Learning Systems
Columbia University
850 Interchurch Center MC 7717
475 Riverside Drive
New York, NY 10115
Phone: 212-870-1289
Fax: 212-870-1285