Semantic processing of Arabic language

Maryam Al-Sadat Hoseini

M.Sc., Department of Arabic literature, Faculty of Literature and Foreign Languages, University of Al-Zahra, Tehran, Iran. m.hoseini1363@yahoo.com

Abstract: In spite of the fact that Arabic offers a well-studied theoretical and historical linguistic knowledge, unfortunately, it has so far received very little computational research and in particular on the level of logical compositional analysis. Furthermore representing Arabic sentences as logic programs has the facility of performing some semantic reasoning tasks on a code based on Arabic predicates. This work is therefore attempting to fill some essential aspects of this gap in introducing a logic-based compositional model covering fundamental issues involved in semantic analysis of Arabic sentences. The focus of attention is relying on studying the compositionality of important Arabic syntactical constituents and on extending the concept of the generalized natural language quantification to Generalized Arabic Quantifiers GAQ utilizing lambda-calculus and the type theoretical analysis of Arabic structure. Since semantic representation has to be compositional in natural language understanding systems this approach attempts to propose an element framework for developing more practical and intelligent Arabic natural language processing systems.

[Maryam Al-Sadat Hoseini. Semantic processing of Arabic language. Journal of American Science 2011;7(4):174-178]. (ISSN: 1545-1003). http://www.americanscience.org.

Keywords: Semantic Processing; Arabic Language; Literature, Formalization, Verbs.

1. Introduction

For the last three decades, concentration on Arabic Language Processing has been focused on the processing of the structure of the language from the morphological and syntactical points of view, whereas research on *computational Semantics* has largely been neglected by Arabic and international computational communities (Mastenbroek, 1994). However, developing natural language understanding systems considering Arabic requires a differentiated and deep semantic processing. This work addresses issues involved in semantic analysis of Arabic and attempts to put fundamentals for the semantic representation in presenting a computational semantic model for Arabic. In the next sections, based on Arabic syntactical constituents utilizing A-calculus and type theoretical analysis of Arabic structure, a semantic model for constructing *meaning representation* of Arabic sentences, will be presented. In addition, this paper is proposing to apply the Generalized Natural Language Quantification concept to *Generalized Arabic Quantifiers, "GAQ"* to capture the specific nature of Arabic semantic compositionality (Beesley, 2001).

2. Literature Review

Semantic processing of human languages is a problematic issue of natural language processing. Artificial Intelligence had a long time ago recognized the importance of semantic representation in context of performing some semantic inferences to achieve human language understanding. Unfortunately, despite the significance of this issue, semantic processing based on logical models in the case of Arabic has so far received very little research attention (Chalabi,2004). Meanwhile, many Arabic morphological analyzers have been successful in solving morphology related issues and many others. Arabic syntax has also been addressed by some researchers, but to some extent and some success has as well been achieved there and others. On the other hand, there were few works reported on the knowledge representation and on the computational semantic of Arabic. Most of the reported works treated this problem informally and from the outside. Semantic analysis and in particular, the problem of the *compositionality* of Arabic has so far not been treated deeply enough, neither linguistically nor logically (Ditters, 2001).

One of the main factors for this negligence might reside in the complexity of this field and in the invisible collaboration between scientists working in the filed of Artificial Intelligence, Arabic, Logic and Linguistics. Therefore, there is a critical need to design sufficient models for semantic processing of Arabic. In spite of the fact, that so far no existing formal theory of semantics is able to provide a complete and consistent account of all phenomena of Arabic and the natural language in general, it remains beneficial to develop models for semantic processing of Arabic even if such models are imperfect or incomplete. Semantic processing has to accomplish different necessary semantic tasks in interrelated and sometimes interchangeable levels to achieve the understanding capability: semantic composition, semantic resolution, and semantic evaluation. *Semantic composition* can be viewed as the process of construction of meaning representation for capturing the *semantic potential of* Arabic sentences. Semantic resolution and semantic evaluation are more concerned with disambiguation under using context knowledge and scoping rules and extracting of relevant information based on performing some deductions and inferences on the semantic representation of a proposition. This work will focus the attention on the fundamentals involved in the *compositionality* of Arabic elementary syntactical constituents and their meaning as a departure point towards developing a potential comprehensive computational semantic for Arabi (Dessouk,1987).

3. Characteristics of the Arabic Language

The Arabic language can be classified into three types: Classical Arabic, Modern Standard Arabic and Colloquial Arabic Dialects. In this paper, we only consider Classical and Modern Standard Arabic and this will be referred to as "the Arabic Language". The Arabic language is composed of nouns, verbs and particles. Nouns and verbs are morphemes and derived from a closed set of around 10,000 roots. Particles are used to complete the meaning of verbs and nouns. The roots are commonly of three or four letters, referred to as triliteral and tetraliteral roots, respectively. Arabic nouns and verbs are derived from roots by applying templates to generate stems and then introducing prefixes and suffixes. It was reported by ElKateb *et al.*, that "85% of Arabic words are derived from triliteral roots". The Arabic verb is any word that indicates the occurrence of an action that is associated with time. An Arabic verb will have a voice (active or passive), a tense (past, present, imperative), a gender (feminine, masculine) and a number (singular, dual, plural). The derivation of the verbs in the different tenses is achieved using well-behaved morphological rules using Eq. (1).

Verb = Prefix1 + Prefix2 + stem + Suffix1 + Suffix2 + Suffix3. (1)

The stem is formed by substituting the characters of the root into certain verb forms, called measures. Arabic verbs can be classified based on the type of the characters forming their root as this will influence their conjugation and the forms of their derivations. Hence, we distinguish two major classes: sound and weak verbs. (Kamp and Reyle, 1993)Sound verbs are verbs whose root does not contain weak letters (i.e. alef ( ), waw (**و**), or yaa ( )); weak verbs are those whose root contains one or more weak letter. The work reported in this paper concerns only derivations from sound verbs.

The measure (also referred to as form or pattern) is defined in as: "a general mould composed of an ordered sequence of characters". There are 37 measures for the triliteral and tetraliteral verbs. Arabic grammarians modeled the formation of nouns and verbs and their derivatives based on the concept of root. This root is a set of the three consonants f 'l **(ف ع ل)** expressing the idea of the action 'to act' (Montague, 1988). For example, the three consonants k t b **(ك ت ب)** expresses the notion of writing and so on. The root is not part of the language; however, to best represent this root Arab grammarians often use the third person masculine in the past tense of a verb. This is similar in meaning to the infinitive mood in English or French languages. The verb kataba (كَتَبَ (To write) is derived from the root "ktb" and scaled to fa'ala فَعَلَ). All verbs have a measure which not only provide morphological information, but in many cases also provide semantic and contextual knowledge. Hence, certain measures can state that the action is performed only once, or performed with some intention etc. Examples showing some of this semantic knowledge will be described later. It is therefore, desirable to define a model to represent the Arabic language that not only models the morphology, but also uses this as the primary source for semantic and contextual knowledge. Hence, in this research, we attempt to use the derivations and their measures to structure the Arabic language and to strongly link the words' morphology to their semantics. This representation is modeled as an ontology. In the following section, we describe the various derivations by providing their measures and then develop the corresponding part of the ontology structure (Black and et al, 2006).

4. Derived verbs

There are two types of verbs in the Ontology, triliteral verbs and tetraliteral verbs. Each triliteral verb will have a set of first stem triliteral derived verbs and a set of first stem tetraliteral verbs.

The first stem triliteral and tetralitera1 verbs are as follow:

Triliteral verbs have the following first stem derivations measures:

(فَعَلَ/ فَعِلَ/ فَعُلَ).

Tetralitera1 verbs have only one first stem measure which is represented as:

فَعْلَلَ.

From these basic forms, many derivatives are produced based on the number of consonants in the verb. The derivation is composed of the basic consonants in the root (three or four characters) to which we add one or more consonants (Friedman-Hill, 2003).

5. Reasons to Choose a Logical Semantic Representation of the Language

There are many reasons to choose a logical language as a target language for the meaning representation. Logic represents a well-known meaning representation formalism that differentiates between syntax and semantics. In addition, it enables inferences over quantified descriptions, which are basic requirements for an adequate meaning representation for any natural language. On the other hand, in spite of the fact that Arabic offers a well-studied theoretical and historical linguistic knowledge, unfortunately, it has so far received very little computational research and in particular on the level of logical compositional analysis. Furthermore representing Arabic sentences as logic programs has the facility of performing some semantic reasoning tasks on a code based on Arabic predicates. Therefore, it is to be expected, that embedding logical formulas with Arabic predicates is a very interesting aspect of logic programming in the context of understanding Arabic. Unfortunately, Arabic NLP researchers have widely neglected this aspect in their published research works.

As Arabic syntax is based on verb-noun in VS, and on *noun-noun* opposition in NS, a semantic correspondence between Arabic sentences and the first order predicate logic, PLl, formulas can be established. The verb as the head of an Arabic Verbal Sentence, and its complements, or the /خبر/; i.e. the nominal predicate as the head of an Arabic Nominal Sentence, can be assigned to a *predicate argument-structure* of the corresponding PLI formula. An Arabic Nominal Sentence can be expressed by using constants or by using quantified arguments of some predicates identifying the role of the subject or the object and other semantic roles.

To interpret logical formulas model theoretically, an *indirect denotation* function is needed to transform higher order logical formulas into PLl. For simplification $[\![\alpha]\!]_{sem}$ is used to denote the *semantic function* of an Arabic syntactical structure "α" such as a feature structure. As this approach is proceeding from the perspective, that Arabic syntactical constituents are able to exhibit relevant compositional *rules* to construct a semantic representation for the most important Arabic sentence structures, the denotation $[\![\alpha]\!]_{sem}$ also has to be *compositional* (Elkateb, 2006).

Logical forms

On the lexical level, an interpretation process might need some conceptual knowledge and some pragmatic contents in form of lexical semantic knowledge or rules to supplement the meaning and to explain the possible word sense potentials of some Arabic natural propositions in a specific domain. For example, interpreting of concepts like some events such (/تعلم/, learn-he-it^{a}) might need some lexical semantic knowledge and pragmatic annotations about their mode, involved objects and their roles, complements, compositional structure and time. This knowledge base can be viewed as kind of a terminology or an ontology describing the involved events and their deep thematic roles including their compositionality encoded in the lexicon. For example, Arabic verbs are *intransitive, transitive,* or di-transitive and therefore, their current argument structure might depend on their contextual interpretation.

6. Definite, Indefinite and Dual

Video The Arabic article (/ال/,The) can be understood as a *determiner.* Determiners are modifiers, which together with nouns or noun phrases build expressions, whose reference can be determined with respect to the referent in a direct way. In the standard analysis of determiners in the type theory an article can be considered as a determiner. Determiners are generally of *type* $\langle\langle e,t\rangle,\langle\langle e,t\rangle,t\rangle\rangle$. Such a type can be expressed using A.-calculus to produce compositional rules for Arabic sentences. In contrast, this view cannot be applied to all Arabic determiner particles directly and in all contexts. The article (/ال/,The) as a logical determiner needs sometimes to be considered in context of some noun phrases. For example, a particle of demonstrative together with the (/ال/) article in (/هذا-ال-كتاب/, this-the-book) can be regarded as a logical determiner (Friedman-Hill, 2003).

7. Formalization of the Language

For the Arabic language to play an important role in this information age, and for the practical applications directly related to the language to be developed to exploit the large amount of information available in resources such as the WWW, there is a need for a proper formalism for the language that is based on the Arabic Language structure and rules governing the formation of its vocabulary. In this section we develop the proposed model which is based first on structuring the Arabic language into a set of equivalent classes and then model each equivalent class as ontology. Hence, a Meta-Ontology that represents the general structures of all these classes is presented (El-Sadany and Hashish, 1989).

8. Logical sentence structures

As mentioned above, Arabic differentiates between different types of sentences:

Verbal Sentences (VS), Nominal Sentences (NS) and *Copulative Sentences.*

On the contrary to European languages, a Verbal Sentence usually starts with a verb, and in most cases has a V-S-O structure. The predicate of a NS usually is a noun, a pronoun, a propositional phrase or an adverb (Gasevic and et. al, 2006). The predicate of VS is a verb and its complements. Copulative sentences have a Nominal Sentence or a Verbal Sentence as a predicate that is bound with the subject through a copulative pronoun.

9. Conclusion

Although, in the Arabic language, triliteral verbs are derived from verbal nouns (مَصْدَر), their complexity, different variations and lack of logical structures makes them extremely difficult to use as the root for deriving verbs. As this study shows, we did find it much easier to derive from verbs as the list of Arabic verbs is known and is finite (countable).

Meanwhile, this work attempted to present some results of a compositional model for logic based semantic representation of Arabic sentences. In this context, this paper has stressed the concept of the Generalized Arabic Quantifiers "GAQ", some potential analysis of state of definite and indefinite in Arabic within different types of Arabic sentences considering the order of words, cardinality, duality, and the meaning of some syntactical constituents. Interestingly, the gathered experiences with this model give strong indications confirming the view that logic based semantic representation for Arabic offers a vital compositionality methodology, which exhibits important logical similarity to the Indo-European languages. As Arabic has received very little computational research on the level of deep semantic analysis, this contribution might encourage some computational linguists and researchers to put more efforts in this complex area of Arabic natural language understanding. In spite of the fact, that so far no existing formal theory of semantics is able to provide a complete and consistent account of all the phenomena of Arabic, it remains beneficial to develop models for semantic processing of Arabic even if such models seem to be incomplete. Currently, I am working on extending this model in considering other semantic phenomena such as resolving some ambiguity and embedding Discourse Representation Theory as a departure point to capture Arabic discourses and features involved in anaphora representations in form of a A-DRT within a Unification based Grammar for Arabic.

Corresponding Author:
Maryam Al-Sadat Hoseini
M.Sc., Department of Arabic literature, Faculty of Literature and Foreign Languages, University of Al-Zahra, Tehran, Iran
E-mail: m.hoseini1363@yahoo.com

References

1. J. Bos, E. Mastenbroek, S. McGlashan, S. Millies and M. Pinkal, A Compositional DRS-based Formalism for NLP Applications, Report 59, VerbMobil, Universitaet des Saarlandes. 1994.
2. K. R. Beesley, Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans 2001, ACL/EACL01, Conference of the European Chapter, Workshop: Arabic Language Processing: Status and Prospects, France. 2001.
3. A. Chalabi, Sakhr Arabic Lexicon, Proceedings of Nemlar International Conference on Arabic Languages Resources and Tools. 2004.
4. E. Ditters, A Formal Grammar for the Description of Sentences Structures in Modern Standard Arabic, A CL/EACL01, Conference of the European Chapter, Workshop: Arabic Language Processing: Status and Prospects, France, 2001.
5. A. EI-Dessouk, Nazif, O. EI-Dessouk and A, Ahmad, An Expert System for Understanding Arabic Sentences, Proceedings of the 10th National Computer Conference, Jeddah, Saudi Arabia. 1987.
6. H. Kamp and U. Reyle, From Discourse to Logic, Kluwer Academic Publishers, Dordrecht. 1993.
7. R. Montague, The Proper Treatment of Quantification in Ordinary English, Philosophy, Language and Artificial Intelligence, eds., J. Kulas, J. H. Fetzer and T. Rankin, Kluwer Academic Publishers, Dordrecht, Boston, London. 1988.
8. W. Black, S. Elkateb, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease and C. Fellbaum, Introducing the Arabic WordNet project, Proceedings of the 3rd Global Wordnet Conference, Jeju Island, Korea, 2006, 22–26.
9. S. Elkateb, W. Black, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease and C. Fellbaum, Building a WordNet for Arabic, Proceedings of The Fifth International Conference on Language Resources and Evaluation. 2006.
10. E. Friedman-Hill, Jess in Action, Manning, Greenwich, UK. 2003.

11. T. A. El-Sadany and M. A. Hashish, An Arabic morphological system, IBM Systems Journal, 28(4), 1989, 600–612.
12. D. Gasevic, D. Djuric and V. Devedzic, Model Driven Architecture and Ontology Development, Springer, Berlin, Heidelberg, 2006.

3/22/2011