Logical Aspects of Unification Based-Arabic Semantic Construction

Bassam Haddad†, Mustafa Yaseen† and Mudhafar Al-Jarrah†

† Faculty of Information Technology
Amman University, Jordan
{haddad, myaseen, maljarrah}@ammanu.edu.jo

Abstract

This paper addresses issues related to employing logic-based semantic construction as meaning representation formalism for Arabic. Since semantic formalism has to be compositional on the level of semantic representation, λ-conversion based on the Discourse Representation Theory can be utilized for realization of semantic construction for Arabic.

Keywords: Logical Form, Arabic Semantic Representation, λ-DRT, Unification-Based Semantics, HDPSG

1. Introduction

For the last two decades concentration on Arabic processing has focused on Arabic from the *morphological* and *syntactical* points of view. In this field, significant progress has been reported (Beesely 2001; Ouersighni 2001; Ditters 2001; Al-Fedaghi and Al-Anzi 1989) and many others.

Despite the importance of *semantic processing* for achieving the *understanding capability*, there was little work reported on *semantic representation* and *semantic analysis* of Arabic (Haddad and Yaseen, 2001; Khayat 1988; Al-Johar and McGregor 1997; El-Dessouki et al. 1988, Al-Muhtaseb and Mellish, 1997) and others. Therefore, we believe that there is an ultimate need to make more effort to develop an *adequate model* for *semantic processing* for Arabic, even though there is no existing *formal theory* capable to provide a complete and consistent account of all phenomena involved in Arabic semantic processing.

Semantic processing has to carry out different tasks on different levels to achieve the *understanding facility*. One of the most important of these levels is the *construction* and *composition* of *meaning representation formalisms* for Arabic sentences. This semantic level plays a decisive role for other semantic processing steps, i.e. *semantic resolution* and *evaluation*.

Lexical and *unification*, especially HDPSG related system development is ongoing in numerous university and industrial settings for different languages. HDPSG is based on GPSG, LFG and Categorial Grammar. In such a grammar the *lexicon* plays a pivotal role, where *semantics* and *syntax* can be integrated in the same grammar. A central concept of a *unification-based grammar* is the notion *"feature"* or *"attribute"* which is characterized *by feature value pairs*.

Applying the operation of *unification* to *two compatible feature values structures* yields a new *feature value structure* containing *all the information* involved in the two original feature structures. The importance of the concept might be residing in the fact that we can solve the problem of finding the right level of *granularity* in classifying words into categories having *internal structures*. Furthermore, the *unification* allows us to combine information from *multiple feature structures*, as long as it is consistent (Pollard and Sag 1994; Bender et al. 1999).

Simulating the *λ-conversion process* in a *feature logical formalism* within a *unification-based grammar* such as HDPSG enables a realization of *unification based semantic construction formalism for Arabic.*

Inspired by the work of (Bos et al. 1994) we propose in this paper to integrate the semantic construction model presented in (Haddad and Yaseen, 2001) in a *unification-based semantic Grammar.*

2. Logical Semantic Representation

Assuring the *modularity constraint* in a natural language understanding system requires a *compositional semantic formalism* on the level of meaning representation. Despite the fact that *predicate logic represents* well-studied formal representation formalism, it does not provide any *compositional facilities*. *λ-abstraction* offers an important framework for achieving such a goal in particular for the meaning construction of Arabic sentences (Haddad and Yaseen 2001; Pinkal 1995; Montague 1974).

In this context we have achieved some success in developing a model for *construction of meaning rep-*

resentation forms for Arabic sentences. Based on some *compositional rules* expressing the meaning of *syntactical categories of Arabic*, our approach employs a *λ-conversion process* to construct *logical forms representing the meaning of Arabic sentences* (Haddad and Yaseen, 2001).

In this model *determiners* play a central role in constructing *semantic constituents*. For example, the Arabic determiners such as "ال_n", "كل", "بعض", etc., could be considered as *quantifiers*. Generally the meaning of a quantifier, ||Quant||, can be expressed as follows:

$$||\text{Quant}|| \Rightarrow \lambda R \lambda S(Quantifier(R,S) \quad (1)$$

The definite determiner "ال_1" combines in general two things together: a *restriction R and a scope S:*

$$||\text{ال}_1|| \Rightarrow \lambda R \lambda S(\text{ال}_1\ (x, R \wedge S)) \quad (2)$$

The following example in *"figure 1, three-branch quantifier tree representation"* might illustrate the basic concept of this approach. Details about this concept are found in [Haddad and Yaseen, 2001].

The *function* of the determiner "ال_1" in the sentence "يتعلم الولد العربيه" can be formulated as follows:

$$||\text{VS}|| \xrightarrow{sem} ||\text{Subj}||\ (||\text{Obj}||\ (||\text{Verb}||)) \quad (3)$$

Applying of (3) to $||\text{ال}_1||$ yields the following logical representation:

$$\lambda R \lambda S(\text{ال}_1\ (x, R \wedge S))\ (||\text{ولد}||)\ (||\text{يتعلم ال عربية}||) \quad (4)$$

$$\lambda S(\text{ال}_1\ (x, \text{ولد}(x) \wedge S))\ (||\text{يتعلم ال عربية}||) \quad (5)$$

$$\text{ال}_1\ (x, \text{ولد}(x) \wedge \text{ال}_1\ (y, \text{عربية}(y) \wedge \text{يتعلم}\ (x, y))) \quad (6)$$

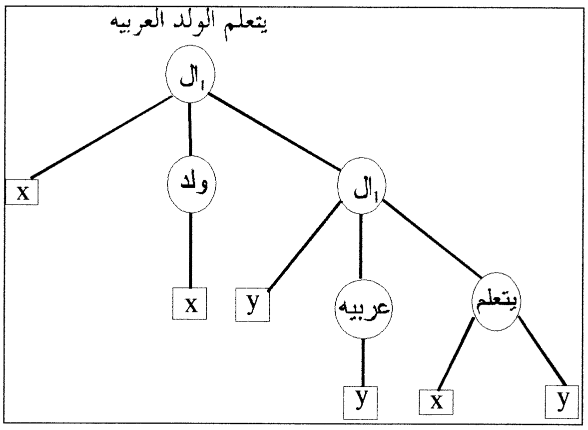


Figure 1: Logical Representation in 3-BQ tree

3. DRT-Based Compositional Aspects for Arabic

Despite the importance of logic-based *compositional models* for achieving Arabic understanding, such methods are rather constructed to deal with Arabic *sentence semantics* and in general they are *inappropriate* for treating *text semantics*.

The *Discourse Representation Theory (DRT)* is capable of capturing problems involved in representing *anaphoric aspects* and *text semantics* (Kamp, 1981; Bende-Farkas and Kamp, 2001).

In this approach the semantic function of sentences consists in constructing of *Discourse Representation Structures (DRS)* by applying certain *DRS construction rules dynamically* within the context of the *referents* in a text.

For instance, the function of a *definite article* seems in the view of *DRT*, not in interpreting it as a *unique quantifier*. It has rather to be understood as a *referent to a certain object in a nominal expression*. Moreover, the interpretation of the *indefinite article* appears *in the first place* not to be as an *existential quantifier*. *An indefinite article introduces* rather a *new referent to the context.*

In addition, one of the most important aspects of *DRT* is its interesting interpretation of *pronouns*. The interpretation of *a pronoun* is not *a variable,* which has to be *locally bound,* but much more as a *definite label* with the function of *making a reference to a previously introduced discourse referent*. Therefore, *a DRT-based semantic construction of Arabic* has to be in the first place not in constructing the logical meaning in an *isolated mode* but much more in a *dynamic and modifiable one.*

Example:

تدرس ماريا لغه تحبها

(Maria studies a language she likes)

The interpretation of this *discourse* starts with an empty *DRS*. After interpretation of the first part of the sentence "تدرس ماريا لغة" (Maria studies a language), the *DRS* is expanded by adding the next *referents* and *conditions*. The referent *e* represents an *event of studying* "تدرس". The *referent n* is used to denote the *time of speech* (see the following figures):

x y e n
"ماريا" *(x)*
لغة*(y)*
e: تدرس*(x,y)*
$n \subseteq e$
DRS for تدرس ماريا لغة

y e n z w s
"تدرس" *(x)*
لغة*(y)*
e: تدرس*(x,y)*
$n \subseteq e$
s: تحب*(z,w)*
$n \subseteq s$
$z = x$
$w = y$
DRS for تدرس ماريا لغة تحبها

In the final stage of representation the resulting *Discourse Representation Structures* are interpreted in *model theory* based logical forms.

It is obvious that *DRT-based semantic construction* proceeds from another point of view than the *Montague-style* in the construction process and it is therefore *not compositional*. Furthermore, the semantic construction is given in *top-down manner* and is *not declarative*, that means the processing order effects the binding possibilities (Pinkal, 1995).

3.1 Compositional Semantics for ARABIC

The Integration of *lambda conversion* in *DRT* extends *DRT* to be *compositional* without losing the important feature of representing *text anaphoric*. In this approach the semantic function of sentences consists in constructing of *Discourse Representation Structures* by applying some *DRS construction rules* within the context of the *referents* in a text. The DRS_n, for instance, consists of a pair: *a universe of discourse*, DR_n, i.e. *a set of Discourse Referents and a set of conditions*, $COND_n$, *about the* DR_n. An additional feature of the language of *λ-DRT, we adopted the merge operation* ⊗, which combines *two DRS's* by taking the *union* of the sets of discourses and conditions separately (Bos et al., 1994):

$$\langle DR_1, COND_1 \rangle \otimes \langle DR_2, COND_2 \rangle \equiv \langle DR_1 \cup DR_2, COND_1 \cup COND_2 \rangle \quad (7)$$

For example the meaning representation for constructing the *DRS* for sentence 'كل طالب يجتهد' could be represented in terms of *λ-DRT as follows (see also (1)):*

$$||\text{كل}|| \Rightarrow \lambda R \lambda S <\{x\},\{\,x: Any\,\}> \otimes R(x) \xrightarrow{\text{كل}} S(x) \quad (8)$$

$$||\text{طالب}|| \Rightarrow \lambda y <\{\},\{y: Individual,\ \text{طالب}(y)\}> \quad (9)$$

$$||\text{يجتهد}|| \Rightarrow \lambda z <\{\},\{e: Event,\ z: Individual,\ \text{يجتهد}(e, z_{<agent>})\}\} \quad (10)$$

The *DRS* in (10) means, that there is an *event* 'يجتهد' which takes an *individual* as an *argument* and plays the role of an *agent*.

Simulating the basic aspects of the *λ-conversion process* presented in [Haddad and Yaseen, 2001] and applying it to the *DRS's* established above would lead in a simplified form to the following semantic representation:

$$<\{x\},\{x: Individual,\ \text{طالب}(x_{<agent>})\}> \xrightarrow{\text{كل}} <\{\},\ \{e: Event,\ \text{يجتهد}(e, x_{<agent>})\}\} \quad (11)$$

3.2 Unification-based Semantic Construction for Arabic

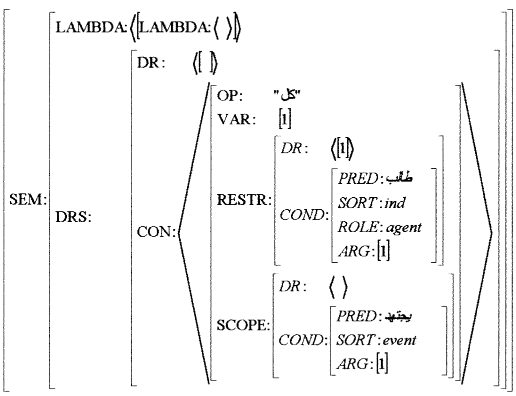
A *λ-Expression* representing the meaning of an *Arabic constituent* (Haddad and Yaseen, 2001) could be formulated in terms of *feature structures*. Such structures might be represented by a LAMBDA and a DRS *feature structure*. A LAMBDA *feature structure* specifies a *list* of the appropriate *arguments, which are* involved in the expression, while *a DRS feature structure represents the body of the λ-expression.* Furthermore, additional *pragmatic notations* could be *also embedded in the DRS feature structures.* Compositional rules expressing the meaning of syntactical constituents are also integrated in the *lexical entries* of a *DRS*.

A *unification-based semantic construction* can be achieved by *unifying* the values of a *LAMBDA feature structure* with *the representations of the feature structures involved in the arguments.* And then storing the results of the unification in the *DRS feature structure of processed syntactical constituent*. This process corresponds to *λ-conversion proposed in* (Bos et al. 1994).

Constructing the meaning of 'كل طالب' in the sentence 'كل طالب يجتهد' requires the application of the *feature structures* involved in (8) to the *feature structures* in (9) (see also "figure 1" and (3), (4), (5), (6)):

$$\left[\text{SEM}: \left[\begin{array}{l} \text{LAMBDA}: \left\langle \left[\begin{array}{l}\text{LAMBDA}: \langle [1] \rangle \\ \text{DRS}: [3]\end{array}\right] \right\rangle \\ \text{DRS}: \left[\begin{array}{l} \text{DR}: \langle\, \rangle \\ \text{CON}: \left\langle \left[\begin{array}{l} \text{OP}: \text{"كل"} \\ \text{VAR}: [1] \\ \text{RESTR}: \left[\begin{array}{l} DR: \langle [1] \rangle \\ COND: \left[\begin{array}{l} PRED: \text{طالب} \\ SORT: ind \\ ROLE: agent \\ ARG: [1] \end{array}\right] \end{array}\right] \\ \text{SCOPE}: [3] \end{array}\right] \right\rangle \end{array}\right] \end{array}\right]\right] \quad (12)$$

To construct the meaning of the whole sentence "كل طالب يجتهد" , "DRS: [3]" has to be applied to the composed DRS in (12):



(13)

It is obvious that (13) corresponds to the logical form in (11).

4. Conclusion

Semantic processing is a non-trivial topic in natural language understanding. We believe that the progress that has been made in recent years is also *applicable to Arabic. Semantic construction* is a substantial task in achieving the basic steps of Arabic understanding. Problems involved in treating *text anaphoric* can be treated *dynamically* by simulating *λ-conversion presented in (Haddad and Yaseen, 2001) within an adapted DRS for Arabic.*

Additionally, this paper is an attempt to direct the attention of research concerned with Arabic processing, in particular to the techniques concerned with *DRT-based and unification-based semantic processing*. At present, our research is concerned with developing such a model by adopting these additional features.

5. References

Al-Fedaghi and Al-Anzi, 1989. Al-Fedaghi, S., Al-Anzi, F. *A New Algorithm to Generate Arabic Root-Pattern Forms*. Proceedings of the 11th National Computer Conference, Saudi Arabia, 1989, pp. 391-400.

AL-Johar and McGregor, 1997. Al-Johar, B., McGregor, J. *A Logical Meaning Representation for Arabic (LMRA)*. Proceedings of the 15th National Computer Conference, Riyadh, Saudi Arabia, 1997, pp. 31-40.

Al-Muhtaseb and Mellish, 1997. Al-Muhtaseb H., Mellish C. *Towards an Arabic Upper Model: A proposal*. Proceeding of the 15th National Conference, Riyadh, Saudi Arabia 1997.

Beesley, 2001. Kenneth 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 , 2001

Bende-Farkas and Kamp, 2001. Ágnes Bende-Farkas and Hans Kamp. *Indefinites and Binding: From Specificity to Incorporation*, Lecture Notes, 13th European Summer School In Logic, Language and Information, ESSLLI, 2001.

Bender, Ivan Sag and Thomas Wasow, 1999. *Instructor's Manual for Syntactic Theory: A Formal Introduction*. CSLI Publications, 1999

Bos, J, E. Mastenbroek, S. McGlashan, S Millies and Pinkal, 1994. *A Compositional DRS-based Formalism for NLP Applications* Report 59, VerbMobil, Universität des Saarlandes 1994.

Ditters, 2001. Everhard Ditters, *A Formal Grammar for the Description of Sentences Structures in Modern Standard Arabic.* ACL/EACL01: Conference of the European Chapter, Workshop: Arabic Language Processing: Status and Prospects, 2001

El-Dessouki, A., Nazif, A, O, Ahmad, 1988. *An Expert System for Understanding Arabic Sentences.* Proceeding of the 10th National Computer Conference, Jeddah, Saudi Arabia, 1988, pp 745-759.

Haddad and Yaseen, 2001. Bassam Haddad and Mustafa Yaseen. *Towards Understanding Arabic: A Logical Approach for Semantic Representation.* ACL/EACL01: Conference of the European Chapter, Workshop: Arabic Language Processing: Status and Prospects, 2001

Kamp, 1981. *A Theory of Truth and Semantics Representation.* In J. Groendijek, T. J. Stokhof, eds., Formal Methods in Study of Languages. Mathematish Centrum, Amsterdam, 1981

Khayat, 1988. Khayat, M.G., Al-Muhtaseb, H.A., *Knowledge Representation in Natural Language System.* Proceedings of the 10th National Computer Conference, Jeddah, Saudi Arabia, 1988, pp. 667-677.

Montague, 1973. Richard Montague. *The Proper Treatment of Quantification in Ordinary English.* In: Philosophy, Language , and Artificial Intelligence, ed., J. Kulas, J. H. Fetzer and T. Rankin, Kluwer Academic Publishers, 1988.

Ouersighni, 2001. Riadh Ouersighni. *A major offshoot of the Dinar-MBC project: AraParse, a morphosyntactic analyzer for unvowelled Arabic* *texts.* ACL/EACL01: Conference of the European Chapter, Workshop: Arabic Language Processing: Status and Prospects, 2001

Pinkal Manfred, 1995. *Sprachverarbeitung: Semantik.* In Einführung in die Künstliche Intelligenz. Ed. Günther Görz. Addison-Wesley, 1995

Pollard and Sag, 1994. Pollard C. and I.A. Sag. *Head-Driven Phrase Structure Grammar.* Stanford, CSLI, 1994.