Extracting Entities and Relationships from Arabic Text for Information System

Belkacem KOUNINEF, Badr AL-JOHAR
Institut National des Télécommunications et des Technologies de l'Information et de la Communication
Route es-Sénia BP 1518, Oran, M'NAOUER 31000, Algéria

Corresponding Author: *bkouninef@ito.dz*

ABSTRACT

This paper addresses the problem of the development of Arabic Natural Language Interface for Information System Design. Most Arabic systems focus on processing Arabic text for machine translation or information retrieval. We present in this paper some aspects for identify Entities and Relationships from a description of the application domain given with a subset of the Arabic Natural Language. The tool starts, in a first step with an interpretation of Arabic text through the system's components (morphological, syntax, semantic analyzers) and generates the meaning representation in a first order logic form. Then in the second step, it uses entities rules and relationships rules for extracting Entities and Relationships, which describes the elements of the Conceptual Schema.

Keywords: *Arabic Text, Morphological, Syntax, Semantic analyzers, Interpretation, Conceptual Schema.*

1. INTRODUCTION

In recent years development of Natural Language Interfaces for Information System (NLIIS) has been one of the most important areas in Natural Language Processing (NLP). We know that it is not easy to develop and implement an Information System (IS) in a company. Information system design is a complicated task. At the origin of any design process, there is observation of the real world.

The nature of the task of design requires an effort of formalization. Our approach is to build a tool, which can be able to carry out the conceptual schema from a description of the application domain given in Arabic text. The goal of the tool is to produce progressively the conceptual schema according an extended Entity-relationship introduced by [4], [5].

In this paper, we focus to understand the mechanisms, which from the observation of the functioning of a real system lead to its representation by Entities and Relationships. In Enterprise the knowledge to the design a part of this knowledge is included in documents (example: forms, documents…) and from which it is possible to deduce a set of entities and a set of rules. The main part-of this knowledge is human knowledge. It is in the managers' brain. These people transmit this knowledge by means of speech more than documents. This is explain why the designer generally uses interviews and pick up this knowledge from speeches. It doesn't exist another way to obtain the complete knowledge about the universe of discourse necessary for design. The initial description of the universe of discourse provided by users of the future information system is made via sentences in Arabic [8].

Our hypothesis is to define a tool for aid to build a conceptual schema for designers in information system design. We have to try to analyze the reasoning process carried out by the designer during the design process. Many semantically rich models are now available [9], [10]. They constitute a contribution at the level of help to information system design, but it is not enough to dispose of pertinent models in order to facilitate the task of design of a complex information system. The models supply concepts by which one can define coherent and readable schemes, but they do not help the designer in the creative work in producing these schemes.

This paper is centered on the step of the design process (figure 1). The formalization of the intellectual process, by which the designer produce a conceptual schema of the information system, from the analyze of the real world, by using Entities rules and Relationships rules. Based on reviews of previous work, there is no such Arabic natural language interface system using this approach [1]. The system proposed in this paper attempts to allow the user to specify universe of discourse with a text written in Arabic.

2. ARCHITECTURE OF THE DESIGN

The system contains several components, which can be grouped in to tow groups. The first group translating the Arabic sentences into a first order logic meaning representation while the second group takes the meaning representation to generate an entity relationship diagram. These components are discussed in the following subsections.

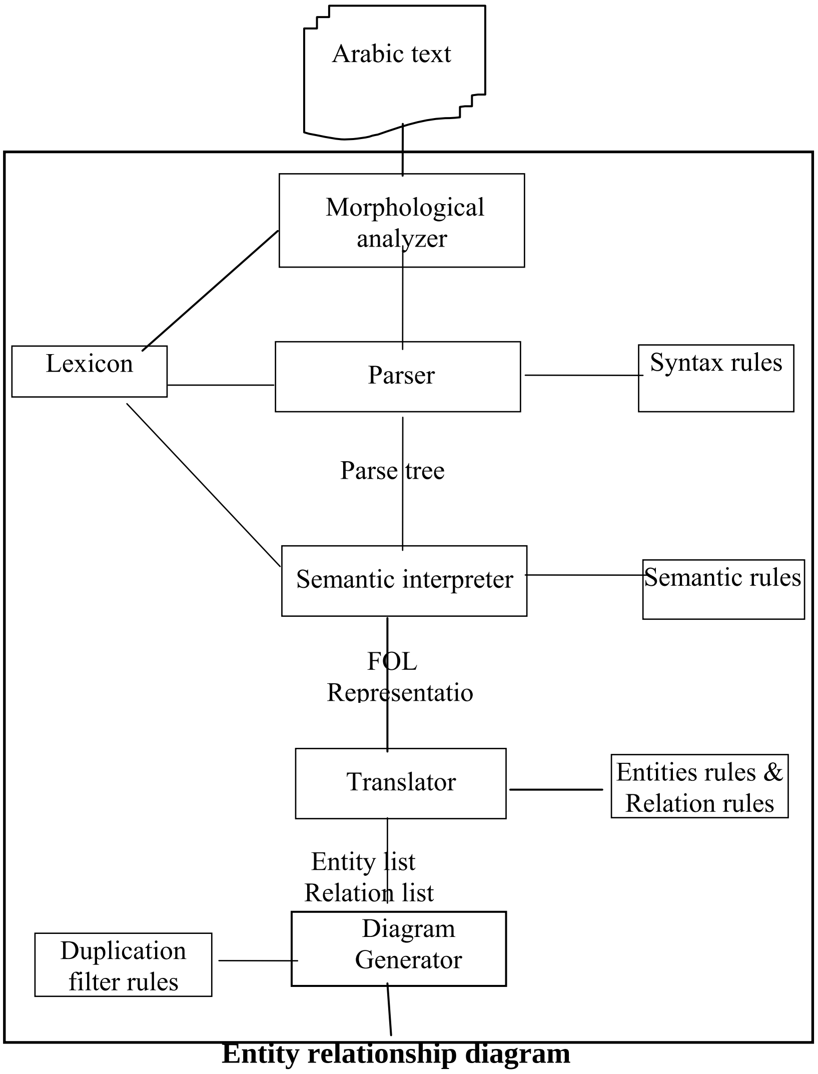

Figure1: Architecture of the system

2.1 Arabic Text

The system can deal with Arabic text describes user specifications. It uses a subset of Arabic language that can be characterized as follows:

1. Sentences must be in present or past tense (مضارع أو ماضي) in the third person of the singular, dual, or plural (masculine or feminine)
2. The use of metaphors is forbidden
3. Sentences must be in the verbal form or nominal form.
4. Sentences that compose the text must be as far as possible independent of each other.

In Arabic, there are three short vowels that correspond to the three cases that occur in Arabic: accusative, nominative, and genitive. These are: - ضمة - Dammah which is a small character above a consonant - فتحة Fatha which is a small diagonal stroke above a consonant, and – كسرة Kasrah which is a small diagonal stroke under a consonant. Also, there is a small notation above a character like this (ّ) called " شدّة - shaddah". This is used when there is a double character, one of them being replaced by this notation. In our system the user must include it in order to get a correct interpretation for the word.

Arabic users rarely use the vowels and most written materials (books, journals, documents, articles, papers) in Arabic do not use these vowels for different reasons. Mainly because it is too time consuming to write each character and its vowel. Therefore, any Arabic natural language processing system should be able to process a sentence without any vowels; and that is what we propose to do.

2.2 Arabic Interpreter

Arabic interpreter contains: a morphological analyzer, syntax and semantic grammar rules, and lexicon. In order to benefit from the tool programs and its Back End, the syntax of both the grammar and the lexicon are based on Generalized Phrase Structure Grammar (GPSG) [3] while the semantic is based on the Property Theory formalism. The following sections give a description of our Arabic interpreter.

2.3 Morphological Analyzer

Words in Arabic are classified into three types: nouns, verbs, and articles. Arabic words can be made up of one or more morphemes; thus, it is very important to extract them. So, the system focuses on the extraction of morphemes from the various inflexions or forms of any word. But not all prefixes, infixes and suffixes are morpheme, sometimes they are just a number of characters to indicate number and gender of the root morpheme of the word. We followed the approach of [6]

For noun lexical entry, the system uses the singular masculine form for regular plural masculine, regular plural feminine, dual masculine, and dual feminine. The irregular plural noun form is used as it is in the lexical entry because most of this form is not based on rules but based on the old Arabic of ten centuries ago. Verbal lexical entries use the singular masculine form in past tense. The following is an example of the system lexical entry:

طالب-talb n.
[agreement,num] = sing.
[agreement,pers] = 3.
[gender] = masc.
[casetype] = human.
[quest] = no.
[nlevel] = bare.
'talb: 'masc.

The system will not keep just the basic stem of the word but also certain other derivations. The reason for that is to keep each word with its semantics. Consider the following words:

سجّل sjl	(verb)	recorded
يُسجل ysjjl	(verb)	records
سِجل sjjl	(noun)	record
مُسجل msjjl	(noun)	recorder
تسجيل tsjjl	(noun)	recording

All of the above words are derived from the basic stem "sjl". Some might ask why not just keep the basic stem of those words and let the morphological analyzer handle the others. The problem with this approach is the loss of the semantics of each word. Verb "sjl = recorded" takes subject and object while verb "ysjjl = records" takes subject and one or two objects and so on for the other words. Further, there is a trade off here between the number of the lexical entries and the number of the morphological rules required.

There are two type of lexicon: static lexicon and dynamic lexicon. The first one contains the basic lexical entries for the system. The morphological analyzer will use this to identify the stems of the word. The dynamic lexicon contains the static lexicon and the new lexical entry built by the morphological analyzer. The dynamic lexicon will be used by the parser to build the syntax and semantics of the sentence.
Each Arabic word is associated with an initial syntactic category and a corresponding semantic value as follows:

"في"	"fy"	prep	λx.λp.λq.'fy(p,q,x)
"مادة"	"madt"	n	'madt
"رسب"	"rcb"	vi	'rcb
"درس"	"drrc"	vt	λp.λq.'drrc(p,q)

The semantic values associated with the nouns and intransitive verbs are simple constants. The semantic value associated with the preposition "fy" is a three place predicate while for the transitive verb it is a two place predicate, etc.

2.4 Parser and Semantic Interpreter

The syntax grammar rules are based on Generalized Phrase Structure Grammar (GPSG) formalism , while the semantic rules are based on the Property Theory (PT) formalism [6]. Each syntax rule is associated with a corresponding semantic rule, in the following way:

$s \rightarrow np , np$ $\| s \| = \| np \| (\| np \|)$
$s \rightarrow np , vp$ $\| s \| = \| np \| (\| vp \|)$
$np \rightarrow det , n$ $\| np \| = \| det \| (\| n \|)$
$vp \rightarrow vi$ $\| vp \| = \| vi \|$

Parser uses the syntax rules to parse the text sentence by sentence and generate parse tree for each one. The first rule represents the nominal sentence without any verb. In this kind of sentence we divided it into two noun phrases: the first one that contains the topic "مبتدأ- almubtada", while the second noun phrase contains the comment "- خبر alkhabar". Its semantic rule states that the semantic value corresponding to an object of syntactic category s is obtained by applying the semantic value of the syntactic category of the first np to that of the second np. The same can be said to the other rules.

Semantic interpreter takes the complete parse tree only and generate the meaning representation for each sentences using semantic rules. Based on the Property Theory, the semantic value associated with "al" is the intentional analogue of a quantifier as:

"الـ"	"al"	det	$\lambda p.\lambda q.\Xi x.px \wedge qx$

2.5 Translator

The translator accepts the meaning representation of the text in First Order Logical (FOL) form. It performs two functions:

1. Extracting Entities by using Entities rules
2. Extracting Relationships by using Relationships rules.

The following examples are the rules for identifying entity and relationship:

E1: If the atom of predicate is a noun
Then take it as an **Entity (e)**

R1: If the atom of predicate is a verb
Then take it as a **Relation (r)**

For each sentence it will generate a list of entities and relations. Consider the following sentences:

S1: الطالب يسجل مواد.
The student registers courses
S2: الطالب يحذف مواد.
The student drops courses
S3: الأستاذ يدرس مواد.
The teacher teaches courses

The semantic interpreter replaces the sentence by the following expression:

∀ x ((طالب student **(x,s)) ⇒**
∃ a (مادة course **(a,b) &** يسجل register **(x,a)))**

∀ x ((طالب student **(x,s)) ⇒**
∃ a (مادة course **(a,b) &** يحذف drop **(x,a)))**

∀ x ((طالب student **(x,s)) ⇒**
∃ a (مادة course **(a,b) &** يدرس teach **(x,a)))**

2.6 Predicates types

When the system recognizes Entities and Relationships, it will represent them in Triplet predicates as follow:

Entity:
e(Num, Atom, Card)
Num: Number for identifying the entity
Atom: String for representing the entity's name
Card: "s" singular or "p" plural.

Relation:
r(Num, Atom, Se(card),De(card))
Num: Number for identifying the association
Atom: String for represent the verb's name
Se: Source entity
De: Destination entity
***Card:* "**s" for singular, "p" for plural.

The predicate participate (e1(card)),r,e2(card)) means that entities e1 and e2 are associated with relationship r. This association should be read as: an entity participates in a relation with certain participation (cardinalities)-constraints.

RS1, RS2, and RS3 represent the Entities and Relationships list for sentences above (S1, S2, S3) accordingly:

RS1: e (1, student طالب, s)
e (2,course مادة, p)
r(1,registerيسجل, e1(s),e2(p))

RS2: e (3,student طالب, s)
e (4,course مادة, p)
r (2,drop يحذف, e3(s), e4(p))

RS3: e (5,teacher أستاذ, s)
e(6,course مادة , p)
r(3,teach يدرس , e5(s), e6(p))

2.7 Diagram generator

As we can see from the example above (RS1, RS2, RS3) the translator may contain a duplication of entities if it mentioned in more than a sentence. Thus we need to apply rules to filter these duplications. D1 is an example of the duplication entities rules:

D1: If the term of predicate X is equal
to the first term of predicate Y
Then let X=Y

After filtering all duplication the tool generate entity relation diagram using Diagram Generator (DG) rules. The following are examples of DG rules:

DG1: If the atom =Entity
Then draw a circle with
Entity's name
DG2: If the atom =Relation
Then draw a line between
two entities
Mentioned and
put the cardinalities.

3. DISCUSSIONS AND CONCLUSION

We have tried to realize a tool for Information System. We have committed an Arabic Front End that takes an input text, producing syntactic and semantic representations, which it then maps into FOL. The syntactic rules are based on GPSG grammar whereas the semantic rules are expressed in Property Theory. The system tested using texts from a well-defined domain. Although, we chose small texts with simple sentences, the system managed to parse and build a meaning representation for up to 86.67% of the given sentences. Most of the failed sentences are because of the lack of semantic rules.

The most important problem is the acquisition and the representation of the knowledge. For the time being we focus on the elements of the Conceptual schema by identifying Entities and Relationships. The system managed to realize entities and relationships of 92.65% from the parsed sentences. This results indicates that once you manage to build formal meaning representation for a given sentences you can easily identify entities and relationship presented on them. The major problem comes from the natural language processing parts (Morphological analyzer, Parser, and Semantic interpreter) rather than the translator part. Our research will be continued to maximize the results and prove this idea using a more complex texts with unstructured domains.

REFERENCES

[1] **Abn-Hesham J :** “mugni allabeeb 3n kutub ala3reeb.” edited by Mazen Al-Mubarrak & Mohammad Ali , Dar AL-Feker, Byroot, 2005

[2] ***Abuleil, S.:*** Extracting names from Arabic text for question-answering systems. In Proceedings of the 7th International Conference on Coupling Approaches, Coupling Media, and Coupling Languages for Information Retrieval (pp. 638–647), University of Avignon (France), 2004.

[3] ***Benajiba, Y and al :*** Arabic named entity recognition: An SVM-based approach. In Proceedings of 2008 Arab International Conference on Information Technology (ACIT) (pp. 16–18). Amman, Jordan

.

[4] ***C. Batini, and all :*** Navathe, “Conceptual Database Design: An Entity Relationship Approach.” The Benjamin/Cummings Publishing Company, Ontario, 1992.

[5] ***Chen P :*** “The Entity-Relationship Model-Toward a Unified View of Data” ACM Transaction on Database Systems, Vol. 1, No. 1, 1976.

[6] ***Feddag A :*** Arabic Morpho-Sytax and Semantic Parsing, In Proceedings of the 3rd International Conference and Exhibition on Multi-lingual computing (Arabic-Roman Script), Cambridge, University,2002.

[7] ***Feddag A and Foxley E :*** “A lexical analyzer for Arabic.” In International Journal of Man-

Machine Studies, Vol. 38, No. 2, PP 189-215 Academic Press, 2008

[8] ***Kouninef B. :*** “Arabic Interface for Information System Design.” Third International conference of modeling & simulation, University of Victoria, 1997, Australia.

[9] ***Shaalan, K., & Raza, H.:*** Person name entity recognition for Arabic, Proceedings of theACL

2007Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources (pp. 17–24). Prague, Czech Republic: Association for Computational Linguistics, 2007.

[10] ***Shaalan, K., & Raza, H.*** Arabic named entity recognition from diverse text types. In GoTAL '08: Proceedings of the 6th international conference on Advances in Natural Language Processing, Vol. 5221 (2008), pp. 440-451