Using MADA+TOKAN to Generate Use Case Models from Arabic User Requirements in a Semi-Automated Approach

Nabil Arman
Department of Computer Science and Engineering
Palestine Polytechnic University
Hebron, Palestine
narman@ppu.edu

Abstract—**Automated software engineering has attracted a large amount of research efforts. The need for new approaches that reduces the cost of developing software systems within project schedule has made it necessary to develop approaches that aid in the construction of different UML models in a semi-automated approach from Arabic textual user requirements. UML use case models represent an essential artifact that provide a perspective of the system under analysis or development. The development of such use case models is very crucial in an object-oriented development methodology. In this paper, MADA+TOKAN is used to parse different statements of the user requirements written in Arabic to obtain different components of a sentence like lists of nouns, noun phrases, verbs, verb phrases, etc. that aid in finding potential actors and use cases. A set of steps that represent our approach for constructing a use case model is presented. Finally, the proposed approach is to be validated and implemented at a later stage of the research project.**

Keywords— Arabic User Requirements, Use Case Model, MADA+TOKAN tool.

I. INTRODUCTION

This Object-oriented methodologies are used for software systems development for the many benefits they provide like software reuse, reducing software development costs, to name just a few. Therefore, there is a need for development of automated tools that can help in constructing different components of an object-oriented software system.

A use case diagram shows a set of use cases and actors and their relationships. Use case diagrams address the static use case view of a system. These diagrams are especially important in organizing and modeling the behaviors of a system. This paper addresses the problem of generating a use case model from user requirements, written in Arabic, in a semi-automated approach. An Arabic natural language processing tool/software, namely MADA+TOKAN, is used to parse different statements of the user requirements, written in Arabic, to obtain lists of nouns, noun phrases, verbs, verb phrases, etc. that aid in finding potential actors and use cases. A set of steps that represent our approach for constructing a use case model is presented.

The rest of the paper is organized as follows: the section about related works presents the literature review and any related approaches; the section about constructing use cases describes the process of constructing use case models from Arabic user requirements; the section about validation presents the validation and implementation of our proposed approach, and finally, the section about conclusion presents the main issues related to the proposed approach.

II. RELATED WORKS

Recently there is a great interest in automating software engineering activities. Many tools were developed to automate different activities of software systems development like normalizing relational database schemas, reverse engineering of relational database and generating the corresponding entity-relationship data model, …etc. [1, 2]. In addition, many CASE tools were developed to aid in drawing different diagrams of UML. For example, Rational Rose is an object-oriented Unified Modeling Language (UML) software design tool intended for visual modeling and component construction of enterprise-level software applications [3]. Rational Unified Process (RUP) is an object-oriented Web-enabled program development methodology. [4].

More advanced tools were developed to automate software engineering activities that are more complicated than just aiding in drawing a UML diagram or checking its overall structure. Arman and Daghameen proposed a systematic approach that generates class diagrams from textual software requirements. They presented some steps to build a matrix that was used to obtain classes and their associations to generate class diagrams [5]. The same authors later developed a CASE tool, called, SDLCCASE tool that implemented their approach [6]. Kothari proposed an approach that can extract the basic elements for generating a class diagram from user requirements written in a clear way. The Natural Language Processing for Class (NLPC) can extract classes, data members and member functions from the given user requirements [7]. This approach was implemented

as a software tool to generate the class diagrams. Seresht and Ormandjieva proposed an approach to generate use case diagrams from software requirements, but this approach depends on other models to obtain the use case by combining two technologies: Recursive Object Model(ROM) and Expert Comparable Contextual (ECC) Models and it doesn't deal with the textual requirements directly [8]. Cayaba et al. proposed an approach called computer automated use case diagram generator (CAUse), that can generate the use case diagrams from a text described using a special language called ADD [9]. However, this approach depends on the ADD language to generate the use case diagrams. Mala and Uma proposed an approach to extract the object-oriented elements of system requirements. This approach started by assigning the parts of speech tags to each word in the given requirements [10]. An automated approach that helps a software engineer in developing formal specifications in VDM is presented in [11]. In this approach, the detection of ambiguous sentences and inconsistences in the informal specifications was a major concern. Relationships are determined from the verbs in the sentences. Entities and relationships are then used to develop an entity-relationship model from which a VDM data types are obtained. Another major research endeavor in automated software engineering was the work of using natural language processing to aid in object-oriented analysis [12]. The natural language processing capabilities to build a UML class diagram was used. The research approach involved two major stages. The first stage is a linguistic analysis of the text to build a semantic net. The second stage uses the semantic net to obtain the class model and its (classes, associations, attributes, etc.). Arman and Jabbarin used Stanford Parser to construct uses cases from Arabic user requirements [13]. In this paper, MADA+TOKAN is used since it provides a richer set of tags that can help in parsing the Arabic statements more accurately. In addition, more accurate heuristics are presented.

III. CONSTRUCTING USE CASE MODELS

The IEEE, in the standard for Software Requirements Specification, identifies a good requirement as correct, unambiguous, verifiable, and traceable. The IEEE also identifies a good set of requirements as complete, consistent, and modifiable. This is an assumption that is used in our approach. It is assumed that the requirements are "good" in the sense implied by the IEEE good requirements assumptions.

This section describes how the actors and use cases are extracted from user requirements written in Arabic. There is a need for an Arabic Natural Language Processing tool such as the MADA+TOKAN, which is used in this research to help in splitting and tokenizing the Arabic user requirements text. Once this is performed, a set of heuristics are used to construct the use case model as presented in subsequent subsections.

A. *MADA+TOKAN*

MADA+TOKAN is a versatile, highly customizable and freely available toolkit for Arabic NLP applications. It consists of two components. MADA is a utility that, given raw Arabic text, adds as much lexical and morphological information as possible by disambiguating in one operation part-of-speech tags, lexemes, diacritizations and full morphological analyses. TOKAN is a utility that, given the information MADA produces, can generate a tokenization (sometimes also called a "segmentation") formatted exactly to user specifications. This tokenization also identifies the stem of the word [14]. All user requirements are processed using the MADA+TOKAN.

A set of user requirements for a system implementing ridesharing is used. The requirements were written in Arabic and some of these requirements are used in our examples. The ridesharing system includes many requirements. Two examples are presented below:

- يقوم السائق بتسجيل الدخول الى النظام ومن ثم يستطيع الاعلان عن الرحله التي سيقوم بها ويقوم في هذه المرحله بتحديد ومتطلباتها وتشمل: وقت الرحله (الانطلاق) و المسار الذي سيسلكه اضافة الى عدد المقاعد الفارغه. كما ويستطيع حذف رحلة بعد انتهاءها او الغائها.

A translation of the examples: "The driver shall be able to sign in to the system and then he shall be able to make an advertisement about the trip he is going to make. At this stage, he provides all information related to the trip, including the time and the number of seats available. He shall also be able to delete the trip afterwards."

- يقوم السائق بقبول الركاب او رفضهم , يستطيع ايضا تتبع المسافرين باستخدام ال GPS ان توفرت هذه الخاصيه عند الركاب في النهايه يقوم بتسجيل الخروج.

A translation of this example: "The driver shall be able to accept or reject passengers. He shall also be able to follow the passengers using a GPS if available. At the end, he shall also be able to sign out."

In addition, MADA+TOKAN uses a set of tags to describe different components of a statement.

MADA+TOKAN tokenizes the statements and uses a large number of tags, including:

MADA+TOKAN has many tags, including, but not limited to:
Verb : VBP, VBZ, VBD
Noun: DTNN , NN ,DTNNS ,NNS ,DTNNP ,NNP ,DTNNPS ,NNPS
Addictive (Object): DTJJ, JJ
Preposition: IN
Connectors: CC (و/ف،أو) , RB ثم

These tags are used in determining the actors and uses cases as described below.

B. *Actors Identification*

To identify the actors from the user requirements written in Arabic, a set of heuristics are presented. These heuristics are used to extract the actors from the tagging of the user requirements generated from the MADA+TOKAN. These heuristics are presented as follows:

- If the statement is simple (i.e. it contains only a verb, a subject and an object) then the actor is the main subject in the statement.

e.g. يقوم السائق بتسجيل الدخول

Using MADA+TOKAN tags, the statement is divided into:

يقوم/VBP السائق/DTNN ب/IN تسجيل/NN الدخول/DTNN

Here the main subject is السائق and it's the actor.

Generalization: If the statement is in the form of <VBP> < DTNN> <IN> <NN> <DTNN> when using the MADA+TOKAN, then the first DTNN is the actor. To simplify referencing, the form can be write as <VBP> < $DTNN_{(1)}$> <IN><$DTNN_{(2)}$> , where the subscripts determine the order of the DTNNs.

- When there are two statements combined with a connection then, there are three cases:

a) The subject is the actor.

e.g. يقوم السائق بتسجيل الدخول إلى النظام و من ثم يستطيع الإعلان عن الرحلة

Using MADA+TOKAN tags, the statement is divided into:

يقوم/VBP السائق/DTNN ب/IN تسجيل/NN الدخول/DTNN يستطيع/VBP ثم/NN من/WP و/CC النظام/DTNN الى/IN الرحلة/DTNN عن/IN الاعلان/DTNN

The actor is السائق.

b) If the subject is redundant in the second statement then the actor doesn't change.

e.g. يقوم السائق بتسجيل الدخول إلى النظام و من ثم يستطيع السائق الإعلان عن الرحلة

Using MADA+TOKAN tags, the statement is divided into:

يقوم/VBP السائق/DTNN ب/IN تسجيل/NN الدخول/DTNN يستطيع/VBP ثم/NN من/WP و/CC النظام/DTNN الى/IN الرحلة/DTNN عن/IN الاعلان/DTNN السائق/DTNN

The actor is السائق.

c) If the subject changes in the second statement then this is another actor.

e.g. يقوم السائق بتسجيل الدخول إلى النظام ومن ثم يستطيع الراكب اختيار الرحلة المعلن عنها من خلال زيارة النظام.

Using MADA+TOKAN tags, the statement is divided into:

يقوم/VBP السائق/DTNN ب/IN تسجيل/NN الدخول/DTNN يستطيع/VBP ثم/RB من/IN و/CC النظام/DTNN الى/IN اختيار/NN الرحلة/DTNN المعلن/DTJJ عن/IN الراكب/DTNN زيارة/NN النظام/DTNN خلال/IN من/PRP ها/PRP

The actors are السائق and الراكب .

Generalization: If the statement is in the form of <$VBP_{(1)}$> <$DTNN_{(1)}$> < $NN_{(1)}$> <$DTNN_{(2)}$> <$IN_{(1)}$> <$DTNN_{(3)}$> <$DTJJ_{(1)}$> <$CC_{(1)}$> <$VBP_{(2)}$> <$DTNN_{(4)}$> <$IN_{(2)}$> <$DTNN_{(5)}$> <$WP_{(1)}$> < $VBD_{(1)}$> <$NNP_{(1)}$> <$CC_{(2)}$> <$NNP_{(2)}$> <$IN_{(3)}$> <$DT_{(1)}$> <$DTNN_{(6)}$> <$NNP_{(3)}$> <$NNP_{(4)}$> <$NNP_{(5)}$> <PUNC> <$NN_{(2)}$> <$JJ_{(1)}$> <$DTNN_{(7)}$> <$DTJJ_{(1)}$> <$CC_{(3)}$> <$DTNN_{(8)}$> <$WP_{(2)}$> <$VBP_{(3)}$> <$N_{(3)}$> <$IN_{(4)}$> <$NN_{(4)}$> <$DTNN_{(9)}$> <$DTJJ_{(2)}$> <PUNC> <$CC_{(4)}$> <$VBD_{(2)}$> <$NN_{(5)}$> <$NN_{(6)}$> <$NN_{(7)}$> <$NN_{(8)}$> <$CC_{(5)}$> <$NN_{(9)}$>

C. Use Cases Identification

To identify the use cases from the user requirements, more heuristics that can be used to extract the use cases from the user requirements are presented.

- If the statement is simple (i.e. it contains only a verb, a subject and an object) then the use case is the main object in the statement.

e.g. يقوم السائق بتسجيل الدخول

Using MADA+TOKAN tags, the statement is divided into:

يقوم/VBP السائق/DTNN ب/IN تسجيل/NN الدخول/DTNN

The main object is تسجيل الدخول and it's the use case.

Generalization: If the statement is in the form of <VBP> < $DTNN_{(1)}$> <IN><$DTNN_{(2)}$>, when using the MADA+TOKAN, then VBP IN $DTNN_{(2)}$ is the use case.

- If the statement contains the connector (و) without any verb or actor in the second statement then the second statement is the use case.

e.g. يستطيع الراكب الانضمام إلى الرحلة و الحجز فيها

Using MADA+TOKAN tags, the statement is divided into:

يستطيع/VBP الراكب/DTNN الانضمام/DTNN الى/IN ها/PRP في/IN الحجز/DTNN و/CC الرحلة/DTNN

In the above example, the statement contains a connector (و) so there are two use cases 1- يستطيع الانضمام 2- يستطيع الحجز

Generalization: If the statement is in the form of <VBP> <$DTNN_{(1)}$> <$DTNN_{(2)}$> <$IN_{(1)}$> <$DTNN_{(3)}$> <CC><$DTNN_{(4)}$><$IN_{(2)}$><$DTNN_{(5)}$> when using the MADA+TOKAN, then

a) The use case is the VBP with $DTNN_{(2)}$.

b) The use case is the VBP with the $DTNN_{(4)}$ after the CC.

- If the statements that contain the connector(أو) without any verb or actor in the statement then the first verb in the statement with the first noun after each connector is a use case.

  e.g. يستطيع الراكب الانضمام إلى الرحلة و الحجز فيها أو الانسحاب منها.

  Using MADA+TOKAN tags, the statement is divided into:

  يستطيع/VBP الراكب/DTNN الانضمام/DTNN الى/IN او/CC ها/PRP في/IN الحجز/DTNN و/CC الرحلة/DTNN من/WP ها/PRP الانسحاب/DTNN

  In the above example, the statement contains a connector (e.g. أو) so there are three use cases 1- يستطيع الانضمام 2- يستطيع الحجز 3-يستطيع الانسحاب.

  Generalization: If the statement is in the form of <VBP> <$DTNN_{(1)}$> <$DTNN_{(2)}$> <$IN_{(1)}$> <$DTNN_{(3)}$> <$CC_{(1)}$> <$DTNN_{(4)}$> <$IN_{(2)}$>< <$PRP_{(1)}$> <$CC_{(2)}$> <$DTNN_{(5)}$> <WP><$PRP_{(2)}$> when using the MADA+TOKAN, then

  a) The use case is the VBP with $DTNN_{(2)}$.

  b) The use case is the VBP with $DTNN_{(4)}$ after the $CC_{(1)}$.

  c) The use case is the VBP with $DTNN_{(5)}$ after the $CC_{(2)}$.

D. Use Case Model Generation

To complete the generation of the use case model, a structure that depicts the relationships among the different tokens is needed. A matrix consisting of columns with headings, which contain the potential use cases, and rows with labels, which contain the potential actors, is used. These are obtained from the heuristics explained previously. The matrix is filled by arrow symbols. An arrow means that an actor is associated with one or more particular use cases. For example, if an arrow is shown in the cell that corresponds to the row labeled with Use Case i and the column labeled with Actor j, t is concluded that Actor j is associated with Use Case i.

Once the matrix is constructed, the use case model is obtained by taking an actor with all its associated use cases to generate a use case diagram. The set of all use case diagrams represent the use case model. According to the above description, this approach can be implemented easily to generate a use case model.

Applying the proposed approach described so far to the set of user requirements mentioned previously generates the matrix presented in Table 1.

As can be seen from the table, all potential actors are associated with the related use cases using the arrow notation.

Table 1. Matrix of Potential Actors and their Use Cases

Potential Actors / Potential Use Case	الراكب	السائق	المدير
تسجيل الدخول	←	←	←
يستطيع الإعلان		←	←
تحديد متطلبات		←	
قبول الركاب		←	
ننبع المسافرين		←	
تسجيل الخروج	←	←	←
يستطيع الانضمام	←		
يستطيع الحجز	←		
يستطيع الانسحاب	←		
يوفر تغذيه	←		
يستطيع اضافة			←
يستطيع حذف			←
تصنيف مستخدمين			←
يستعرض الرحلات	←		←

IV. Proposed Approach Validation and Implementation

The next step in this research is to validate the proposed approach. Once the approach proves to be beneficial, the proposed approach will be implemented as a software tool that can be used to generate the use case model from Arabic user requirements.

V. Conclusions

The proposed approach of developing use case models is very essential in the practice of object-oriented software engineering. This approach can be implemented and incorporated in any Integrated CASE (Computer Aided Software engineering) Tool to aid in the process of obtaining the use case models from user requirements written in Arabic. The approach has the main advantage of dealing with Arabic language. In addition, a set of heuristics are presented to obtain the use cases. These heuristics use the tokens produced by a natural language processing tool, namely MADA+TOKAN. These tokens are then used as the main components of the use case diagram, namely, the actors and the use cases. Finally, the proposed approach is to be validated and implemented in further research efforts.

Acknowledgment

The author would like to than the Software Engineering Research Group members at Palestine Polytechnic University, especially Mr. Ibrahim Nassar for his help regarding MADA+TOKAN tool and Dr. Dia Abu Zeineh for his help regarding the use of Natural Languages Processing tools.

REFERENCES

[1] N. Arman, "Normalizer: A Case Tool to Normalize Relational Database Schemas, Information Technology Journal, pp. 329-331, Vol. 5, No. 2, ISSN: 1812-5638, 2006.

[2] N. Arman, "Towards E-CASE Tools for Software Engineering," International Journal of Advanced Corporate Learning, pp. 16-19, Vol. 6, No. 1, 2013.

[3] http://searchciomidmarket.techtarget.com/home/0,289692,sid183,00.html, accessed: October 15, 2013.

[4] "Rational Unified Process (RUP)": ch1, Prentice Hall 1990, ISBN 0-13-629841-9.

[5] N. Arman. and K. Daghameen, "A Systematic Approach for Constructing Static Class Diagrams from Software Requirements," International Arab Conference on Information Technology (ACIT2007), November 26-28 2007, Amman, Jordan.

[6] K. Daghameen and N. Arman. "Requirements Based Static Class Diagram Constructor (SCDC) CASE TOOL." Journal of Theoretical & Applied Information Technology, pp. 108-114, Vo15, No. 2, 2010.

[7] P. Kothari, "Processing Natural Language Requirement to Extract Basic Elements of a Class," International Journal of Applied Information Systems (IJAIS) , ISSN : 2249-0868.

[8] S. Seresht and O. Ormandjieva, "Automated Assistance for Use Cases Elicitation from User Requirements Text," 11th. Workshop on Requirement Engineering, 2009.

[9] C. Cayaba, J. Rodil and N. Lim, "CAUse: Computer Automated Use Case Diagram Generator", 2006.

[10] G. Mala and G. Uma, "Automatic Construction of Object Oriented Design Models [UML Diagrams] from Natural Language Requirements Specification", 2006.

[11] F. Meziane, "From English to Formal Specication",PhD thesis Dept. of Maths and Computer Science, University of Salford, UK, 1994.

[12] H. Harmain, "Building Object-Oriented Conceptual Models Using Natural Language Techniques", PhD thesis, University of Sheffield, 2000.

[13] N. Arman and S. Jabbarin, "Generating Use Case Models from Arabic User Requirements in a Semiautomated Approach Using a Natural Language Processing Tool," Jouranl of Intelligent Systems, Vol. 24, No. 3, 2015 (to appear).

[14] N. Habash, O. Rambow, and R. Roth, "MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In Proceedings of the Second International Conference on Arabic Language Resources and Tools", pp 102-109, Cairo, Egypt, 2009.