The Enhanced Arabchat: An Arabic Conversational Agent

Mohammad Hijjawi
Faculty of Information Technology
Applied Science Private University
Amman, Jordan

Zuhair Bandar
School of Computing
Manchester Metropolitan University
Manchester, UK

Keeley Crockett
School of Computing
Manchester Metropolitan University
Manchester, UK

***Abstract*—The Enhanced ArabChat is a complement of the previous version of ArabChat. This paper details an enhancements development of a novel and practical Conversational Agent for the Arabic language called the "Enhanced ArabChat". A conversational Agent is a computer program that attempts to simulate conversations between machine and human. Some of lessons was learned by evaluating the previous work of ArabChat . These lessons revealed that two major issues affected the ArabChat's performance negatively. Firstly, the need for a technique to distinguish between question and non-question utterances to reply with a more suitable response depending on the utterance's type (question and non-question based utterances). Secondly, the need for a technique to handle an utterance targeting many topics that require firing many rules at the same time. Therefore, in this paper, the "Enhanced ArabChat" will cover these enhancements to improve the ArabChat's performance. A real experiment has been done in Applied Science University in Jordan as an information point advisor for their native Arabic students to evaluate the Enhanced ArabChat.**

Keyword—Artificial Intelligence; Conversational Agents and Arabic

I. Introduction

From Turing test (imitation game) time[1], which he was tried to solve his test by answering his question "if a computer could think, how could we tell?", number of researches tries to solve his test by developing a conversational agent. A Conversational Agent(CA) is a computer program that attempts to simulate conversations between machine and human [2]. Since that time, number of CAs types has been raised due to the diversity of applications that's could CAs applied in. This is including Embodied CA, Linguistic CA and mixed approach between them [3].

The Embodied type has a humanoid or animated character which shows a body reactions such as facial expressions, eyes movement and the character sounds [3]. Linguistic CAs deals with spoken or/and written conversations without to embed the embodied abilities. Finally, the mixed approach which can share the features of both types [3].

This paper is interested to build a linguistic CA. Therefore, the main approaches that be used to build such types of CAs will be introduced which are Natural Language Processing (NLP), Semantic Sentence Similarity (SSS) measures and Pattern Matching (PM) [4].

The NLP which is defined in computing as "the computational processing of textual materials in natural human languages" [5] is based on understanding a sentence. Technically, NLP-based CAs uses grammar rules and a list of attribute/value pairs to extract the conversation's speech act type from the sentence [6]. Then, it uses these extracted information to fill a template-based response [6]. However, extraction such information is not easy at all as it depends on many linguistic factors [6]. In a rich language especially the sematic languages such as Arabic, this extraction will be harder to process [4, 6].

The SSS approach is based on checking the similarity level in semantic between two sentences [3]; the first sentence is the conversation itself and the second is a scripted pattern inside the CA. The most closed pattern in semantic (meaning), its response will be replied as an answer to the conversation. The SSS approach is based on computational semantic based manual built databases such as WordNet [3]. However, such database established in 2006 [3], and the research in SSS, in general, is still a young research area in the Arabic language [4].

The PM approach is the most common used approach for its simplicity and because it is language independent [4]. It does not need complex pre-processing stages like the previous approaches, so it is not expensive computationally. Consequently, a number of CAs such as [4, 7-9] used this approach to handle conversations for applications deal with large numbers of users in a real-time environment like the Internet [10]. Basically, this approach based on matching a conversation with a pre-structured patterns to find the suitable one. Then, the response that related to the best matched pattern will be replied [4]

All the three approaches have advantages and disadvantages that can be cleared in different references such as [3, 4, 8, 11-14]. However, as mentioned above, this paper is considered as a complement work for the previous edition of ArabChat [4] to enhance its performance and called the "Enhanced ArabChat". The rest of this paper describes the enhancements details. The next sections describe the Enhanced ArabChat framework and the conducted experiment and its results. Finally, a general evaluation has been conducted to evaluate the proposed CA.

II. The Enhanced ArabChat

The previous work of ArabChat [4] considered the first phase of ArabChat development. In this paper, number of novel modules has been integrated into the ArabChat to lunch the second phase of it to be the "Enhanced ArabChat".

Before proceeding with the enhancement explanation, it is important to summarise the first edition of ArabChat framework. The ArabChat is PM approach which based on pattern matching technique to handle the Arabic textual user conversations.

The ArabChat is a rule-based CA modelled into three main modules which are scripting language, engine and brain [4]. The scripting language is a predefined language used to script an application's domain in order to represent it. The scripting language is structured as a rule-based language that contains contexts (main domain topics) which each context has several rules (sub-domain topic) and each rule has number of patterns (the simulated user sentence) and responses. While, the brain is a structured knowledge base that is used to store the domain's scripts. Finally, the engine handles user's utterances (conversations) by matching them with the scripted domain and replying with a suitable response. The conversation remains ongoing until one of the conversation's parties (user and ArabChat) terminates it. The ArabChat has the right to terminate the conversation and close a session for many reasons described in [4]

As discussed before, number of lessons have been learned by evaluating the previous work of ArabChat [4]. These lessons revealed that two major issues affected the ArabChat performance negatively. Firstly, the need for a technique to distinguish between question and non question utterances in order to reply with a more suitable response depending on the utterance's type. Secondly, the need for a technique to handle an utterance targeting many topics that require firing many rules at the same time. For instance, ArabChat has two rules to deal with two different topics which are "Accommodation" and "Transportation" in Jordan, and the user targets the two topics in the same utterance like "how much is the cost for the student accommodation in Jordan and how much is the average cost for the transportation as well". ArabChat was unable to reply for both topics (the rule that has the best matched pattern will be fired). Therefore, these issues have been taken into consideration in order to improve ArabChat's performance and continue developing to generate the Enhanced ArabChat. These two issues are related to the ArabChat engine's work but also they need amendments on the rule's structure (scripting language and brain) in order to meet these engine-based improvements. Therefore, it can be summarised that all the new required amendments can be classified as engine-based amendments and scripting language-based amendments.

The engine-based amendments deal with developing the two required modules. Firstly, the need for a technique to distinguish between question and non question utterances. Therefore, in the Enhanced ArabChat, the module "Utterance Classification" which deals with this issue will be developed. The second module is to handle an utterance requiring the firing of many rules at the same time. Therefore, the module "Hybrid Rule" which deals with this issue will be developed in the Enhanced ArabChat.

The scripting language-based and brain amendments summarised by the need of adding some new features to the rule's structure in order to meet the requirements of the new amendments of the Enhanced ArabChat engine which are the "Utterance Classification" module and the "Hybrid Rule" module. In addition, other features of the rule's structure will be added in order to facilitate evaluating the Enhanced ArabChat.

III. The Enhanced ArabChat Framework

The Enhanced ArabChat framework is a complement of the first version of ArabChat framework. Consequently, The Enhanced ArabChat includes all of the developed modules in the first version of ArabChat and the new integrated developed modules ("Utterance Classification" and "Hybrid Rule") as Figure 1 describe. The new two developed modules (("Utterance Classification" and "Hybrid Rule") had been published in two different papers in [15, 16] that related to the same research work for building the Enhanced ArabChat.

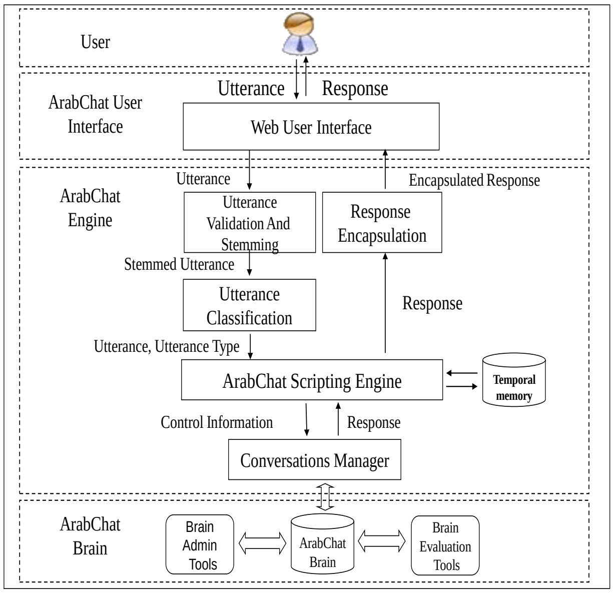


Fig. 1. The Framework of the Enhanced ArabChat

A. *The Enhanced ArabChat Scripting Language*

The scripting language of the Enhanced ArabChat is a complement to the scripting language of the previous version of ArabChat [4]. The new modules in the Enhanced ArabChat require amending the rule's structure. Therefore, only the new amendments on the rule's structure will be described in this section. In the Enhanced ArabChat, a rule has three new parameters, which are:

- "نوع القاعدة" "Rule type".
- "تتعارض مع" "Conflict With".
- "المعلومات المطلوبة" "Information Requirements".

The first parameter ("Rule Type") has been added due to the new module "Utterance Classification". This module classifies the utterance into question and non question utterances. The question utterance requires firing question-based rule. While, the non question utterance requires firing non question-based rule. As a result, each rule in the Enhanced ArabChat has its own type either a question-based or non question-based rule.

The second parameter ("Conflict With") has been added due to the new module "Hybrid Rule". This module deals with an utterance requesting different information that requires firing many rules. In reality, some rules (topics) may be conflicting with each other. These conflicting rules should not be fired together for the same utterance. If so, the rule that has the highest strength will fire. The parameter "Conflict With" is used to alert the engine that the rule has raised a conflict with other rules, which prevents the engine from firing them together.

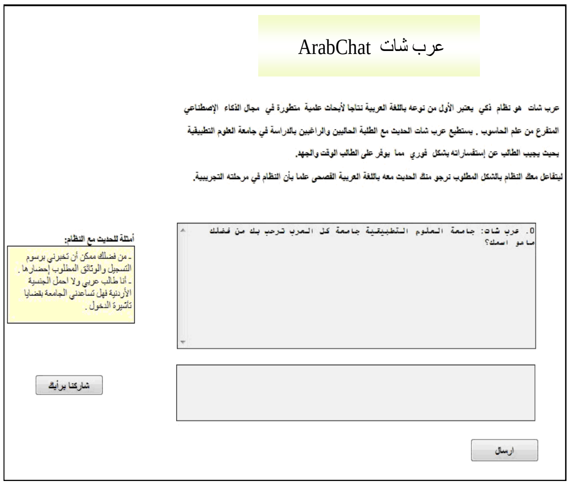
The third parameter ("Information Requirements") has been added to the rule's structure for the new ArabChat evaluation purposes. A rule's "Information Requirements" is a collection of words that should exist in an utterance in order to fire the rule. This parameter will be described in details later.

B. The Enhanced ArabChat framework components

In this section, only the new modules in the Enhanced ArabChat will be described. The new module "Utterance Classification" embedded in the Enhanced ArabChat engine where the module "Hybrid Rule" embedded in the Enhanced ArabChat scripting engine. In addition, a little amendment has been done on the web user's interface in order to let users to evaluate the Enhanced ArabChat by filling a questionnaire as described in the next section.

1) Web User Interface

The Web User Interface (WUI) for the Enhanced ArabChat has the same functions as the previous version of ArabChat WUI. In addition, in this WUI (for the Enhanced ArabChat), an online user questionnaire is added and linked by the button "شاركنا برأيك" "share your opinion" as shown in Figure 2. This questionnaire will be used to evaluate the Enhanced ArabChat. Moreover, there is a box to show the user examples of how they can communicate with the system.


عرب شات ArabChat

عرب شات هو نظام ذكي يعتبر الأول من نوعه باللغة العربية نتاجا لأبحاث علمية متطورة في مجال الذكاء الإصطناعي
المتفرع من علم الحاسوب . يستطيع عرب شات الحديث مع الطلبة الحاليين والراغبين بالدراسة في جامعة العلوم التطبيقية
بحيث يجيب الطالب عن إستفساراته بشكل فوري مما يوفر على الطالب الوقت والجهد.
ليتفاعل معك النظام بالشكل المطلوب نرجو منك الحديث معه باللغة العربية الفصحى علما بأن النظام في مرحلته التجريبية.

أمثلة للحديث مع النظام:
- من فضلك ممكن أن تخبرني برسوم التسجيل والوثائق المطلوب إحضارها .
- أنا طالب عربي ولا احمل الجنسية الأردنية فهل تساعدني الجامعة بقضايا تأشيرة الدخول .

0. عرب شات: جامعة العلوم التطبيقية جامعة كل العرب ترحب بك من فضلك ما هو اسمك؟

Fig. 2. The Enhanced ArabChat WUI

2) The Enhanced ArabChat engine

The Enhanced ArabChat has a new integrated two modules which are "Utterance Classification" and "Hybrid Rule". As discussed above, only the new modules in the Enhanced ArabChat will be described in this paper.

a) Utterance classification

If the validation process detects a valid utterance, then the utterance classification process will start. The Enhanced ArabChat will first classify the utterance as either a question or non question utterance. This classification is based on a set of rules generated from a top-down decision tree induction technique as described in [15].

The "Utterance Classification" module is fully described in a related research work for ArabChat in [15]. In this section, the need of classifying an utterance will be only described. Consider the following rule "How-travel-to-Amman" which is designed to answer utterances asking about the ways of travelling to Amman:

< How-travel-to-Amman >

a: 0.1

p: * الذهاب إلى عمان *   * go to Amman *

p: الطرق * إلى عمان *   * ways * to Amman

p: الذهاب إلى عمان *   * go * Amman

r: بإمكانك الذهاب إلى عمان من خلال الحافلات العمومية المتواجدة في مجمع الحافلات

(You can go to Amman by public buses that stationed in the buses station)

Consider the two utterances " كيف أستطيع الذهاب إلى عمان من الزرقاء؟" "How I can get to Amman from Zarqa?" and " أنا أفضل الذهاب إلى عمان في الصيف" "I like to visit Amman in the summer". The two utterances matched the same pattern " * الذهاب إلى عمان *" because they shared the same keywords as the pattern. However, the first utterance is considered as question while the second is considered as non question. The two different utterances require different responses. Therefore, scripting two different types of rules (question and non question) for the same topic is important to deal with question and non question utterances. However, scripting the two rules will not solve the problem properly as their patterns may still share the same keywords. Consequently, the rule that has the highest strength (best matched) will fire [4]. Therefore, adding extra keywords to the question based rules' patterns is important. These extra keywords might be some interrogative words such as "كيف" "How". For instance, the pattern might be "كيف * الذهاب إلى عمان *" in order to match the mentioned question utterance.

Scripting all of the expected interrogative words for each question-based rule will increase the number of patterns. Alternatively, the "Utterance Classification" module has been developed. In addition, this module might increase the Enhanced ArabChat performance by replying to an utterance depending on its type (question or non question). For a specific topic, a question-based utterance might need an accurate answer for the question, while a non-question-based utterance might need agreement or disagreement with the user's thoughts.

The "Utterance Classification" methodology itself is already explained in details in a related research work for ArabChat in [15]. However, in this paper [15], a novel technique has been proposed and developed to classify the Arabic sentences into questions and non-questions based sentences. This classification was based on structural information contained in Arabic function words. The developed technique extracts the function words features by replacing them with numeric tokens and replacing each word with a standard numeric token. The extraction process or the classification rules based on building a decision tree and it provides a high effective classification results.

After determining the utterance's type, the utterance and its type are sent to the scripting engine in order to deal with the utterance depending on its type.

b) The Enhanced ArabChat scripting engine

The "Enhanced ArabChat" scripting engine is a complement of the previous version of ArabChat scripting engine. The integration of the "Utterance Classification" module will affect the methodology of the Enhanced ArabChat scripting engine methodology. After integration, the Enhanced ArabChat scripting engine works depending on the classified utterance's type as depicted in Figure 3. When the utterance is classified as question, the engine explores the question-based rules as depicted in Figure 4. Contrarily, if the utterance is non-question, the engine deals with non question-based rules. The Figure 4 represents the scripting engine methodology of the Enhanced ArabChat after classifying the utterance.

When the utterance is classified as a question but the engine cannot match it with any question-based rules, it will keep the generated utterance's type as it is. Then, it will switch to explore the non-question-based rules (as Figure 3 displays). This switching process (in case no matching occurs in question-based rules) has been adopted to give the utterance another chance to match non question-based rules. If there is no matching in the non-question-based rules as well, the engine checks the previous processed context's patterns if the previous processed utterance's type is the same type as the current processed utterance's type. If so, the scripting engine starts matching the previous context's patterns with the processed utterance. If matching occurs, the scripting engine checks if the utterance tries to target many rules (the utterance has many requested information that require firing many rules, see the "Hybrid Rule" as depicted in Figure 4.

Usually, the user's utterance has one topic (one topic means the utterance targets one rule) to deal with. Other utterances might have many topics to deal with which requires firing many rules for the same utterance. The Enhanced ArabChat scripting engine was designed and developed to deal with this issue.

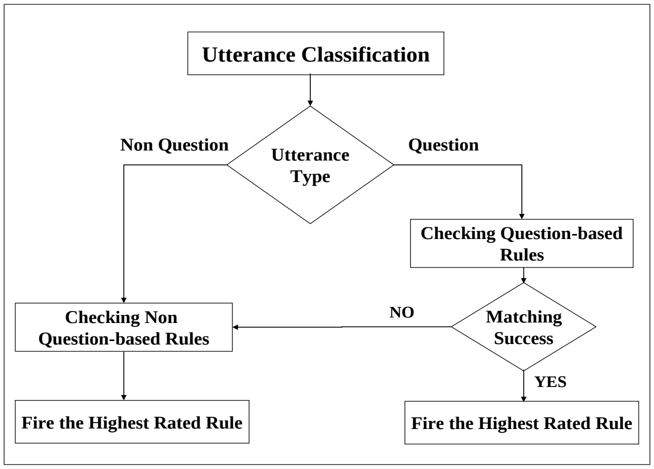

Utterance Classification
Non Question
Utterance Type
Question
Checking Question-based Rules
Checking Non Question-based Rules
NO
Matching Success
YES
Fire the Highest Rated Rule
Fire the Highest Rated Rule

Fig. 3. The Enhanced ArabChat scripting engine methodology based on utterance type

When the scripting engine detects that the utterance has multiple requested topics that require firing many rules, it will deal with all of them in one rule known as the "Hybrid Rule". For instance, an utterance is requesting information about the documents that are needed to register in a university and the fees of registration. Assuming the Enhanced ArabChat has two different rules to deal with these different topics (registration's documents and registration fees). If so (the utterance has the two topics), the scripting engine will deal with them by start creating the "Hybrid Rule" in temporal memory as described later.

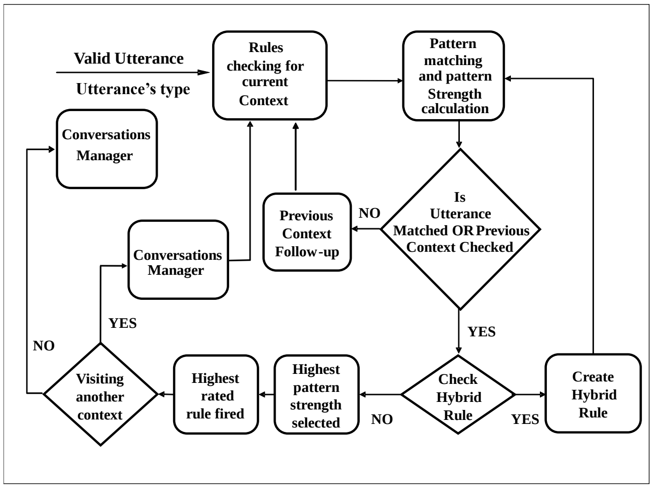

Valid Utterance
Utterance's type
Rules checking for current Context
Pattern matching and pattern Strength calculation
Conversations Manager
Is Utterance Matched OR Previous Context Checked
NO
Previous Context Follow-up
Conversations Manager
YES
NO
YES
Visiting another context
Highest rated rule fired
Highest pattern strength selected
NO
Check Hybrid Rule
YES
Create Hybrid Rule

Fig. 4. The Enhanced ArabChat scripting engine methodology

The Hybrid Rule

Before developing the "Hybrid Rule", when two or more patterns belonging to different rules match against a user's utterance, only the rule that has the highest pattern strength will fire. Matching different patterns for different rules means that the utterance contains (requests) different topics, and the ArabChat replies with just one topic which could be considered a weak point in any CA.

Therefore, the “Hybrid Rule” has been proposed and developed in the Enhanced ArabChat in order reply to an utterance requesting a number of topics.

A “Hybrid Rule” is a hybridization process used by Enhanced ArabChat scripting engine to hybridize all rules that their patterns matched the processed utterance in one rule called the hybridized rule or “Hybrid Rule”. After this hybridization, the “Hybrid Rule” will have the ability to let the engine to reply to all targeted topics in the utterance.

When the Enhanced ArabChat scripting engine detects that many patterns belong to different rules match the same utterance, it starts the hybridization process for the matched rules by accumulating their information (as described in [16]) and creates the “Hybrid Rule” in a temporal memory. A Hybrid Rule’s structure is like any rule’s structure in terms of having an activation level and all other components. However, a Hybrid Rule differs from other rules with the number of patterns and responses. A “Hybrid Rule” will only have one hybridized pattern to match with the utterance that contains many targeting topics and one hybridized response to reply to the targeting topics.

Sometimes a user might merge a large number of topics inside the same utterance, which could lead to generating a very long response. Therefore, in order to avoid a very long response, the Enhanced ArabChat enables the scripter to determine the maximum number of rules to be fired for the same utterance, depending on the highest rules’ strengths priority. Some rules might conflict with each other if they are already designed to deal with opposite topics. When a user targets such conflicting rules in the same utterance, the engine will not fire these conflicting rules for the same utterance in terms of naturalness response. Instead, the Enhanced ArabChat will fire the rule that has the highest strength among them.

As discussed earlier in this section, the Hybrid Rule was created in temporary memory and is kept until the next user’s utterance is processed, to ensure that the consecutive utterance does not target the same topics that the Hybrid Rule handled. If so, the Enhanced ArabChat will re-send the Hybrid Rule’s response to the user in order to avoid recreating the Hybrid rule and thus reducing the processing time.

IV. EXPERIMENTS AND EVALUATION

As discussed earlier in this paper and in the published related work in [4], there are number of lessons that have been learned by analysing the first version of ArabChat logs. As a result, these lessons led to continue development to generate the Enhanced ArabChat. In this paper, a number of experiments will be conducted in order to test the full Enhanced ArabChat. In addition, the evaluation methodology (RMUT) that was adopted to evaluate the first version of ArabChat [4] does not give a precise result. Therefore, a new comprehensive evaluation methodology which gives more precise results will be conducted in this paper to evaluate the Enhanced ArabChat. The Enhanced ArabChat evaluation methodology is comprised of two main approaches: namely, objective approach and subjective approach. The objective approach will be applied through developed automatic evaluation measures and logs. The subjective approach will be performed with recourse to human judgment using the user’s questionnaire.

The Enhanced ArabChat’s applied domain is the same domain as the first version of ArabChat. However, the improvements that were conducted in the Enhanced ArabChat’s engine led to modify the Enhanced ArabChat applied domain scripts’ structure (contexts and rules) to meet the new modifications. Consequently, the Enhanced ArabChat’s applied domain contexts and rules were structured to consist of both question-based and non question-based contexts and rules.

Table 1 represents the first version of the ArabChat applied domain’s contexts. In the first version of ArabChat, the contexts numbers are from 1 to 5 (see Table 1) and their rules are regarded as non question-based in the Enhanced ArabChat scripts because they deal with non question-based topics. Context numbers 6 to 32, apart from 30 and 31 (the contexts numbers 30 and 31 regarded as non question-based rules), may be targeted by question-based and non-question-based topics. These contexts (from 6 to 32 apart from 30 and 31) and their rules are regarded as question-based in the Enhanced ArabChat in order to deal with question-based utterances as their scripts were already scripted to deal with this type of utterances. Figure 5 represents the first version of ArabChat applied domain diagram. Then, in the Enhanced ArabChat, a new 25 contexts and their rules have been scripted and regarded as non question-based in order to deal with non question-based topics. Subsequently, patterns scripting process is done for these rules (non question-based rules) in order to match non question-based utterances. In total, the Enhanced ArabChat applied domain consists of 57 contexts, 907 rules and 20944 patterns as depicted in Figure 5.

TABLE I. THE FIRST VERSION OF ARABCHAT APPLIED DOMAIN'S CONTEXTS

#	Context Name	Context Name in English
1	البداية	Home Context
2	اسم المستخدم	User Name Capturing
3	فحص الكتابة بغير اللغة العربية	Test input language
4	الخروج	User exit (session terminating)
5	ألفاظ سيئة	Bad words
6	معلومات عامه عن الأردن	General Information about Jordan
7	معلومات عامه عن جامعة العلوم التطبيقية	General Information about ASU
8	المواصلات	Transportation
9	السكن	Accommodation
10	الإقامه /تأشيرة الدخول	Visa/residency
11	القبول والتسجيل	Acceptance and registration
12	عدد الساعات المعتمده للتخصص	Courses credit hours
13	رسوم الساعه المعتمده للتخصص	Courses hour's fee
14	الحد الأدنى لمعدل القبول في التخصص	Minimum average of a course acceptance
15	التحويل	Transfer
16	التأجيل	Postpone
17	دليل الهاتف	Phone book
18	دليل البريد الالكتروني	E-mail book
19	دفع الرسوم	Fees paying
20	خصومات الطلبة	Fees discounts
21	الدورات التدريبية	Training courses
22	التأمين الصحي	Health insurance
23	التجسير	Bridging
24	توفر التخصص والكلية	Courses availability
25	خدمات الجامعه	University services
26	الدراسات العليا	Postgraduate studies
27	المرشد الأكاديمي للطلبه	Academic advisors
28	متابعة الطلبة الخريجين	Graduated students' follow-up
29	أسئلة شخصية قد تسأل للنظام	Questions to ArabChat
30	لمتابعة الحديث	To continue conversations
31	معالجة الجمل الناقصة(تخصصات)	Dealing with utterances with insufficient explanation
32	مدن الأردن	Jordan cities

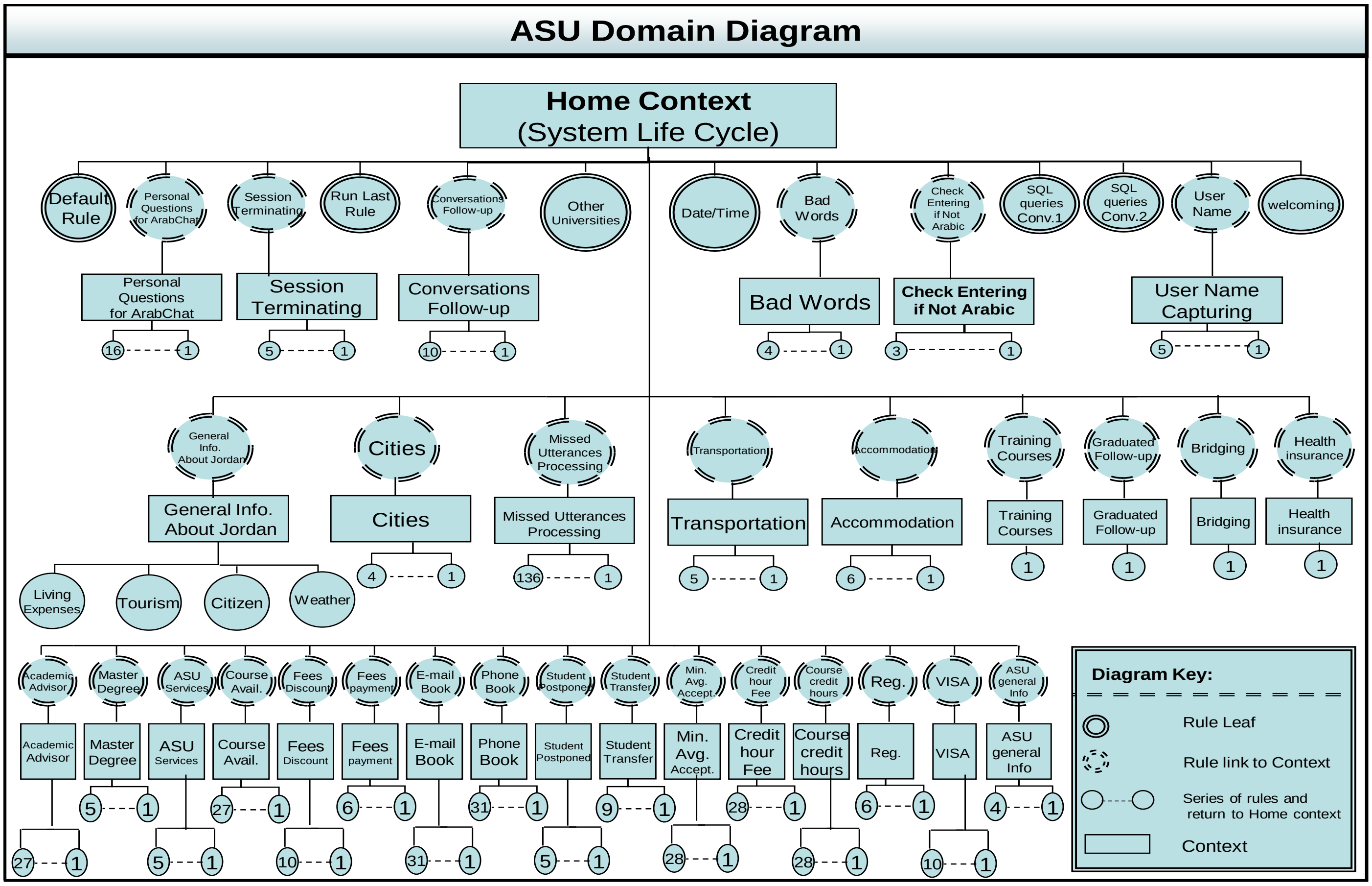


Fig. 5. The first version of ArabChat applied domain diagram

A. Experiment 1

Experiment 1 was conducted to test the full Enhanced ArabChat capabilities from different aspects. Firstly, the experiment tested the Enhanced ArabChat's scripting engine in terms of its ability to recognise patterns' wildcards and matching utterances properly. In addition, test the "Hybrid rule" feature in the scripting engine. Secondly, experiment 1 examined the Enhanced ArabChat classification methodology (the Enhanced ArabChat classifier). Moreover, through analysing the un-matched utterances log, the experiment investigated the applied domain if it is meet the users' needs.

Experiment 1 methodology

The Enhanced ArabChat was deployed on the ASU (Applied Science University) website [17] and accessed by all qualified users such as registered students, non registered students, and employees. The Enhanced ArabChat was available online and in use for 23 days.

Experiment 1 results

The Enhanced ArabChat handled 1766 utterances from 203 users, an average of 8.699 utterances per user. The most accessed contexts are presented in Table 2; these contexts were reported through using the automatic "Domain statistics" report (see [4]).

Table 3 represents the targeting distribution for experiment 1 users based on core domain contexts and general domain contexts. The core domain refers to contexts related to ASU students' issues, while the general domain represents the remaining contexts. Table 3 contents was manually collected through classifying the domain's contexts into "General Domain" and "Core Domain" contexts and then summation the number of targeting for each class's contexts (number of targeting for a context can be reported using the automatic "Domain statistics" report, see [4]) after classifying it whether it is related to core domain or general domain. The results of the Enhanced ArabChat classifier are presented in Table 4. In this table, the number of classified utterances, divided based upon utterance type are presented. All of experiment 1's results will be discussed with the evaluation of the Enhanced ArabChat in the next section.

TABLE II. THE MOST 5 TARGETED CONTEXTS

#	Context Name	Times been targeted (Percentage)
1	Courses Fees (Question-based)	239 (13.533 %)
2	Admission/Registration (Question-based)	193 (10.928 %)
3	Bad words (Non question-based)	146 (8.267 %)
4	Continuing Conversation (Non question-based)	103 (5.832 %)
5	Accommodation (Question-based)	61 (3.454 %)
Total		**742 (42.015 %)**

Utterance classified Type	Number of classified utterances (Percentage)
Question-based Utterance	1005 (56.908%)
Non Question-based Utterance	761 (43.07%)
Total	(100%)

TABLE III. CORE DOMAIN VS. GENERAL DOMAIN; DISTRIBUTION TARGETING BY USERS

Scripted Domain Type	Times been targeted (Percentage)
General domain	302 (25.209 %)
Core domain (Information Point Adviser)	896 (74.791 %)
Total	**1198 (100 %)**

TABLE IV. QUESTION VS. NON-QUESTION CLASSIFICATION RESULTS BY ENHANCED ARABCHAT

The Enhanced ArabChat evaluation based on experiment 1 results

The Enhanced ArabChat will be evaluated using a comprehensive evaluation methodology depending on the results of experiment 1. This evaluation methodology aims to check whether the components of the Enhanced ArabChat are doing their tasks properly ("Glass box approach"). Moreover, the opinions of the Enhanced ArabChat users in experiment 1 will be taken into consideration in this evaluation methodology.

Therefore, this evaluation methodology consists of two main parts, which are objective and subjective evaluations, to cover all previously mentioned aims. The objective evaluation will be used to examine the Enhanced ArabChat as one component and the Enhanced ArabChat individual components using automatic techniques or logs manual checking, as will be discussed in the next section. On the other hand, the subjective evaluation will be done with recourse to users' judgment. Therefore, a questionnaire has been developed in order to ask the users about their opinion of using the Enhanced ArabChat.

1) The Enhanced ArabChat objective evaluation

The objective evaluation has been done based on the "Glass box approach". The "Glass box approach" evaluates the main components of the Enhanced ArabChat, namely, the scripting engine, the "Hybrid Rule", the applied domain coverage, and the "Utterance Classifier". In addition, a manual check will be conducted to examine the Enhanced ArabChat interaction speed performance and to determine the most Arabic interrogative used in users' utterances.

a) The Enhanced ArabChat scripting engine Evaluation Aim

The evaluation aim is to test the functionality of the main component of the Enhanced ArabChat: the scripting engine. This evaluation will determine whether or not the scripting engine is doing its tasks properly such as recognising patterns' wildcard, matching utterances successfully and navigates among the scripted contexts.

Evaluation Methodology

This evaluation will be done by determining the RMUT (Ratio of Matched Utterances to the Total) of the Enhanced ArabChat users. The RMUT (see [4]) is automatically calculated per user by the Enhanced ArabChat once a user session is closed and it can give a general overview of scripting engine's performance.

Evaluation results and discussion

The evaluation results show that the average RMUT for the 203 users of the Enhanced ArabChat is 67.836%. According to Table 2, the third most targeted context by users was the "Bad words" context which means existing of unserious users and by checking the "Unmatched" log, it has been noticed that number of utterances were unmatched due targeting this context with uncovered keywords or colloquial words. However, by checking the "Unmatched" log, it is possible to notice that the unmatched utterances were due to the use of colloquial words such as using the colloquial phrase "شلونك" instead of "كيف حالك" "How are you". In addition, there were a number of unmatched utterances due to misspelled keywords such as "مندق" instead of "فندق" "Hotel". Moreover, there were a number of unmatched utterances due to missing patterns or missing keywords in the scripting patterns. Finally, it has been noted that the fewer amount of unmatched utterances was due to targeting uncovered topics such as requesting more accurate information about tourism in Jordan or asking about courses teachers names, which is outside the main scope of the Enhanced ArabChat domain. Given this, it is possible to conclude that the Enhanced ArabChat scripting engine achieved a reasonable performance (67.836%.of all utterances) in terms of its ability to handle conversations successfully.

b) The "Hybrid Rule" module evaluation

Evaluation aim

The aim is to evaluate the Enhanced ArabChat scripting engine performance in terms of its ability to fire more than one rule for the same utterance at the request of that utterance.

Evaluation methodology

The evaluation methodology is based on manually analysing the Enhanced ArabChat logs in order to determine the utterances that targets more than one topic in them which requires firing more than one rule. Then, analysing the Enhanced ArabChat responses for such utterances and determining manually whether or not the number of replied topics has been performed.

Evaluation results and discussion

Analysing the Enhanced ArabChat logs revealed that there are 121 utterances targeting more than one topic. From those utterances, 85 utterances targeting two topics for the same utterance such as “ ما هو سعر الساعة لتخصصي المحاسبة ونظم المعلومات الادارية” “How much is the credited hour for accounting and management information systems courses” and “ ما سعر تخصص التمريض وكم عدد ساعاته” “How much is the nursing course and how many credited hours it have”. In the first utterance, the user targeted two different rules in the same context, which are “Accounting fees” and “Management Information systems fees”, while the second utterance targeted two different rules related to two different contexts, which are “Nursing fees” and “Nursing credit hours”. In contrast, 36 utterances targeted 3 topics for the same utterance.

For utterances that targeted two topics, the Enhanced ArabChat replied successfully to 82.354% of them (70 utterances) with the two targeted topics, while the Enhanced ArabChat replied successfully to 7.058% of utterances (6 utterances) with one topic only. The Enhanced ArabChat failed to reply to 10.588% of those (9 utterances) with any topic. Instead, it fired a default rule for the current processed context. The reasons for un-replying to some topics was due to either missing patterns or topics being outside the domain, such as “متى يبدأ التسجيل في الجامعة وما اسم مدير التسجيل” “When the registration in the university will begin and what is the registration manager name”. In this utterance the second part of it asking about the registration department manager’s name is not covered in the scripted domain. Consequently, the Enhanced ArabChat will reply to only the first part of the utterance asking about the first date of registration, which it already covered in the scripted domain. For utterances that targeted three topics, the Enhanced ArabChat replied successfully to 75% of them (27 utterances) with the three targeted topics. In contrast, the Enhanced ArabChat replied successfully to 13.888% of those (5 utterances) including two topics only. The Enhanced ArabChat failed to reply to 11.111% of those (4 utterances) with any topic. The reasons of un-replying to some topics caused by either missing patterns or topics outside the scripted domain. Given this, the “Hybrid Rule” implementation performance is good enough to fire more than one rule for the same utterance, which means replying to more than one topic at the same time. This might increase the Enhanced ArabChat performance.

c) Domain coverage evaluation

Evaluation aim

The aim is to evaluate aspects related to the scripted domain. Firstly, whether or not the scripted domain coverage was adequate and covered the user needs. This metric will show if the scripted contexts and rules are sufficient to answer all ASU students concerns. Secondly, the evaluation will discuss the most targeted contexts reported in Table 2. The most targeted contexts were measured to identify the topics that users show most interest in and then giving them more priority for future scripting. Thirdly, the evaluation will discuss the users’ targeting distribution for the core domain contexts and the general domain contexts, which is reported in Table 3. This will help finding which type of domain contexts the user is most interested in.

Evaluation methodology

The scripted domain coverage was determined manually by checking the content of the “Unmatched” log (see [4]). This log contains all unmatched utterances by the engine. These were mainly because; either the unmatched utterance’s topic was not covered in the domain, or the topic is covered, but there is no pattern that matched the utterance. For each context, the Enhanced ArabChat automatically accumulates the number of times it is targeted using the “Domain Statistics” tool (see [4]). Given this, the most targeted contexts are easy to report. The targeting distribution by users between the core domain contexts (contexts numbers from 6 to 57 apart from 29 and 30) and the general domain contexts (contexts numbers from 2 to 5) was manually determined by counting the number of times each context has been targeted, which was already reported using the “Domain Statistics” tool [4].

Evaluation results and discussion

As discussed earlier, it has been revealed that the fewer amounts of unmatched utterances was due to uncovered targeted topics, such as requesting more accurate information about tourism in Jordan or asking about courses teachers names. In addition, some users converse about more detailed topics such as fees, discounts, and student transferral issues require asking the ASU employees. All of these uncovered topics are outside the main scope of the Enhanced ArabChat domain. As a result, Enhanced ArabChat domain coverage is enough to handle ASU’s students’ utterances. The most accessed contexts are presented in Table 2 and Table 3 presents the accessed distribution by users between the core domain and the general domain. According to Table 2, three of the five most highly accessed contexts are related to the core domain, namely course fees, admission/registration, and accommodation. Although most contexts were students related, there were general contexts in the Enhanced ArabChat which addressed other issues such as “general information about Jordan”. Most users concentrated on issues related to their concerns (student issues) rather than general issues. This was expected because the Enhanced ArabChat was employed as an information point advisor for ASU students. The second highly accessed context is “Admission/Registration”. This context deals with issues related to admission, registration fees and procedures, indicating that a large number of users were unregistered students. The third most targeted context is “Bad words” that includes rude utterances containing impolite words. This context, which has a high accessed rate, is negatively affecting the results because many of these utterances went unmatched and affected the RMUT ratio. The fourth most accessed context is “Continuing Conversation”. This context deals with utterances usually used by users in order to continue the conversation, such as “ok”, “great”, and “that’s fine”. Table 3 shows results that confirm that ArabChat’s users’ focus was on contexts related to their concerns (student issues) rather than those dealing with general issues not directly related to the ASU. As mentioned earlier, this is expected as the Enhanced ArabChat was employed as an information point advisor for ASU students.

As a result, the coverage of general issues (general domain contexts) is reasonable for the nature of work of the Enhanced ArabChat as the most interested domain is the core one.

d) The Utterance Classification module Evaluation

Evaluation Aim

The evaluation aim is to test the performance of the “Utterance Classification” module in the Enhanced ArabChat. As discussed earlier, the “Utterance Classification” module was developed to classify the processed utterances into question-based and non-question-based utterances.

Evaluation Methodology

The Enhanced ArabChat classifies the utterance and stores its type (question-based or non-question-based) in the “Brief log” [4]. A manual checking of the classified utterances in the “Brief log” was conducted in order to find out the real correct classification accuracy.

Evaluation Results

The Enhanced ArabChat “Utterance Classification” module results are presented in TABLE IV. , while the results of the manual check of the classified utterances are presented in Table 5.

TABLE V. ACCURATE (QUESTION VS. NON-QUESTION) CLASSIFICATION RESULTS (MANUAL CHECKING)

Utterance classified Type	Number of classified utterances (Percentage)
Question-based Utterance	1312 (74.292%)
Non Question-based Utterance	454 (25.708%)
Total	(100%)

Discussion

Table 4 shows that the Enhanced ArabChat users entered 56.908% of their utterances as question-based utterances. The manual checking of the logged utterances revealed the real percentage of the question-based utterances is 74.292% and not 56.908% as the “Utterance Classification” module generated. The manual checking showed that misclassified question-based utterances were due to the following reasons:

- Conversing using Colloquial Arabic, containing stop words written the Colloquial way such as using “شو” instead of “ماذا” “What” and “هسا” instead of “الان” “Now”. The adopted stop words list by the Enhanced ArabChat’s classifier does not contain such words as they differ among people and thus it is very hard to enumerate them.

- Attaching some interrogative words with other words not applicable in Arabic such as “ ماتكلفة المواصلات باليوم الواحد” “How much is the transportation per day”. The interrogative word “ما” “How much” should not be attached to any word.

- Constructing questions semantically such as “ أعلمنا عن الطقس في عمان” “Tell us about the weather in Amman”. Although, the Enhanced ArabChat’s classifier was trained on indirect questions, but questioning phrases like “أعلمنا” “tell us” “قل لي” “tell me” does not involve the stop words list because they are not stop words. In addition, stop words can accept affixes which might lead to the generation of new words and mislead the classifier.

- Using the interrogative word “أ” “Alef” to build the question such as “أأنت هو المسؤول هنا” “Are you the responsible here?”. The Enhanced ArabChat’s classifier does not learn on these instances, and the stop words list does not contain such an interrogative word (“أ”) because it does not appear alone in Arabic.

- Using stop words in the questions not covered in the “Utterance Classification” stop list such as “فبماذا” “so using which” and “ولماذا” “and why”. The interrogative word “ماذا” “What” accepted the prefix “فب” to generate the stop word “فبماذا” “so using which”. When interrogative words or other stop words accept affixes, enumerating and detecting them becomes impossible.

The misclassified non question-based utterances were due to the following reasons:

- Conversing using Colloquial Arabic, which makes it very hard to detect stop words.

- Using interrogative words in their utterances constructed to seem like questions but in reality, were used for purposes other than questioning, such as “ ما أجمل جامعتكم” (“How beautiful your university”) and “ ما أحد حضر إلى الدورة إلا أنا” “No body came to the course except me”.

The reported number of the entered question-based utterances might be considered large. This is may be due to various factors that need further investigation. These main factors are presented in Table 6 and labelled depending on the potentially responsible party including user, engine, the selected domain, and scripts.

TABLE VI. FACTORS OF THE HIGH PERCENTAGE OF QUESTION-BASED UTTERANCES THAT NEEDS VERIFICATION

#	Factor description	Caused by
1	It might have occurred due to the nature of the applied domain (Information Point Advisor) which is expected to deal with question-based utterances rather than non question-based utterances.	The selected domain
2	It might have occurred as Enhanced ArabChat scripting engine does not deal properly with non question-based utterances.	The engine
3	It might have been caused by the lack of knowledge and experience of users in terms of the nature of CAs and they are dealing with it as QA system.	The user
4	It might have occurred because the Enhanced ArabChat responses do not encourage people to continue conversations based on non-question utterances. Consequently, the users keep asking questions to ArabChat.	The scripts

In Table 6, there are 4 main factors that might have caused the high percentage of question-based utterances. The nature of the applied domain might have encouraged people to find their requested information rather than participate in normal chatting as it works as information point advisor for issues of students. The results in Table 3 show that 74.791% of utterances accessed the core domain (information point advisor) which supports the first factor and leads to accept it as one of the reasons that caused this large number of question-based utterances.

The second factor (caused by the engine) might need further investigation through another experiment in order to investigate it by conversing with the Enhanced ArabChat using non question-based utterances and monitoring the outcomes. Hence, experiment 2 (discussed later) will deal with this issue. The third factor related to the user will be investigated in the Enhanced ArabChat subjective evaluation (discussed later).

Analysing the Enhanced ArabChat logs has revealed that scripted responses might have failed to encourage users to engage in general conversation with the agent through non-question-based utterances. Unfortunately, as depicted in Figure 5, the applied domain is quite large, which led to increases the difficulties of scripting the large amount of rules. Their responses should have a crafted response to try guide the user indirectly and encouraging him/her to keep conversations going. Accordingly, all of these reasons lead to accept the fourth factor which emphasizes that the responses scripts were one of the reasons for the large amount of question-based utterances.

e) The most used Arabic interrogative words in utterances

Evaluation aim

This evaluation aims to determine the most used Arabic interrogative words in the users' utterances. Determining these words will help in discovering the most used types of questions, such as questions regarding quantities, places, time, or yes/no questions. Then, enhancing the Enhanced ArabChat classifier with the ability of classifying such types of questions.

Evaluation methodology

A manual identification for all question-based utterances from the "Brief Log" has been conducted. Then, a manual counting of the Arabic interrogative words has been done individually.

Evaluation results

The used Arabic interrogative words in users' utterances are presented in Table 7. This table presents the most used Arabic interrogative words and their number of usage.

TABLE VII. MOST USED ARABIC INTERROGATIVE WORDS

#	Interrogative word	Interrogative word Count
1	كم (How much)	289
2	ما, ماذا (What)	243
3	كيف (How)	201
4	هل (Is, are)	127
5	لماذا (Why)	93

Discussion

According to Table 7, users are most interested in questions about issues related to fees, quantities, and manners. This was obvious from the first three interrogative words ("How much", and "what"). In Arabic, the interrogative word "ما" ("what") could be used to ask about fee such as " ما سعر هذا الكرسي؟" ("How much is this chair?") or to ask about quantity such as "ما عدد الطلاب في الجامعة؟" "How many students in the university?".

Other question types users were interested in include "Yes/No" questions, which was conducted based upon from the fourth most used interrogative word. While the last most used interrogative word is related to questions about reasons.

As mentioned earlier, determining the most used interrogative words might help in discovering the most used types of questions and then improving the Enhanced ArabChat classifier's capabilities and thus increasing the quality of Enhanced ArabChat response. In order to do this, further research is needed to classify the question-based utterances into other categories such as questions about places, times, people, and yes/no questions. Then, Enhanced ArabChat tested the compatibility between a question type and a response in terms of whether or not the response met the question's type. If no compatibility is found, other research work would need to be done to develop new techniques to deal with this issue.

f) The Enhanced ArabChat interaction speed evaluation

Evaluation aim

This evaluation aims to check Enhanced ArabChat's interaction speed. Interaction speed refers to the time that Enhanced ArabChat takes to reply to a user. This speed might be used to evaluate the usability of Enhanced ArabChat.

Evaluation Methodology

The Enhanced ArabChat stores the elapsed time that it is taken to process each utterance in the "Brief log" [4]. Then, a manual classification for all utterances into valid and invalid utterances was conducted. In addition, another classification has been conducted to categorise the valid utterances based on the number of targeted rules, one rule or many rules (Hybrid Rule). Finally, a manual calculation for the average of elapsed time for all utterances that related to the previous categories was calculated individually.

Evaluation results

The following results were achieved:

1. The general average elapsed time to process an utterance that access one rule is 1.52 seconds.
2. The average elapsed time that is needed to process an utterance that accesses number of rules (Hybrid Rule) is 3.24 seconds.
3. The average elapsed time is needed to process an invalid utterance is 0.6 seconds.
4. The general average elapsed time to process all utterances is 1.869 seconds.

Discussion

Results showed that the time needed by Enhanced ArabChat to process an utterance is based on the status of the processed utterance (valid or invalid). The invalid utterance needs an average of 0.6 seconds to process. The elapsed time that need to process a valid utterance based on the number of rules that need to be fired in order to handle that valid utterance. For instance, the utterance that requires firing one rule needs less time than utterance requires firing many rules.

The reported average of elapsed time for all processed utterances is 1.869 seconds. This amount of time might be considered a good result, especially as the Enhanced ArabChat handled utterances through the Internet.

Summary of the Enhanced ArabChat components evaluation based on the results of the objective (Glass box) approach

Through evaluating the Enhanced ArabChat using the “Glass box” approach, the following outcomes have been found and they are summarised as follows:

- The evaluation of Enhanced ArabChat scripting engine results show that the average RMUT for the 203 users is 67.836%. This result should be better but unfortunately it has been affected negatively by the number of unserious users as discussed earlier. In addition, by analysing the “Unmatched Utterances” log, it has been revealed that all of the unmatched utterances were due number of reasons as mentioned earlier but not due to an engine failure. Given this, it might be considered that the Enhanced ArabChat scripting engine achieved a reasonable performance in terms of its ability to handle conversations successfully.

- The Enhanced ArabChat scripting engine dealt successfully with utterances targeting many topics which requires firing many rules at the same time. The scripting engine replied successfully to 82.354% of utterances targeting two topics where it replied successfully to 75% of utterances targeting three topics at the same time. Consequently, the “Hybrid Rule” feature in the scripting engine that dealt with these kinds of utterances, performed successfully.

- The Enhanced ArabChat classifier achieved a reasonable performance, which demonstrates that the “Utterance Classification” module had a good methodology of classifying utterances. However, it has been noticed that there are large numbers of question-based utterances. As discussed earlier, the nature of the applied domain (information point advisor) and some of the Enhanced ArabChat response styles were two factors that caused this large number of question-based utterances. However, two other factors, the user and the engine need to be verified will be discussed later in this section.

- As discussed earlier, it has been revealed that the fewer amount of unmatched utterances was due targeting uncovered topics. The rest of unmatched utterances were due to the use of colloquial keywords, misspelled words and missing patterns. The uncovered topics that caused a number of unmatched utterances were outside the main scope of the Enhanced ArabChat domain. As a result, the Enhanced ArabChat domain coverage is enough to enable the Enhanced ArabChat to work as an information point advisor and handle users’ conversations successfully.

- According to tables Table 2 and Table 3, the Enhanced ArabChat users (ASU students) were more interested in conversing about issues related to them (core domain) rather than general topics (general domain). In addition, Table 7 confirms that the Enhanced ArabChat users were more interested in asking about quantities, fees and manners.

- Finally, it has been noticed that the elapsed time that Enhanced ArabChat needs to process an utterance depends on the number of rules that needed to be fired to handle the utterances. The general average of the elapsed time to process all utterances is 1.869 seconds. This is can be considered a good result, especially considering that the Enhanced ArabChat works in an online environment (the Internet).

The next section describes the second part of the evaluation which is the subjective evaluation.

2) The Enhanced ArabChat subjective evaluation

The subjective evaluation aim

As discussed earlier, the subjective evaluation will be conducted by asking the Enhanced ArabChat experiment 1 users to give their opinion about various aspects of using the Enhanced ArabChat. Therefore, an online questionnaire was developed and placed on the same web user interface used to converse with the Enhanced ArabChat. The subjective evaluation (the online questionnaire) aims to enable users to evaluate the Enhanced ArabChat user interface, usability, naturalness, the applied domain coverage, speed, availability of Similar Arabic agent, and user general satisfaction.

The subjective evaluation methodology

The online questionnaire has 14 questions designed to meet the above mentioned evaluation aims. For each aim, a number of questions have been assigned to determine the user opinions concerning them. For each question in the questionnaire, a user has 3 options from which to select his/her degree of approval or disapproval for the asking issue. These options are “موافق” (“Agree”), “محايد” (“Neutral”), “ غير موافق” (“Disagree”). The following are the questionnaire questions (14 questions):

1. “واجهة النظام كانت مناسبة جدا” “The user interface was suitable”.
2. “كان النظام قادر على إجابتك على جميع إستفساراتك” “The agent was able to answer all your utterances”.
3. “أجوبة النظام كانت واضحة ومفهومة.” “The agent responses were clear and understandable”.
4. “لم تواجهك أية مشاكل فنية عند إستخدامك النظام” “You experienced no technical problems whilst using the agent”.
5. “الوقت المستغرق من النظام للرد على استفساراتك كان مناسبا” “The elapsed time taken by the agent was reasonable”.

6. “ تفاعل النظام معك كان واقعي وحقيقي شبيه بتفاعل الانسان من حيث الأجوبة وردود الفعل” “The interaction with the agent was realistic and believable”.
7. “ صعوبة التخاطب مع الجامعة عبر الهاتف والبريد الالكتروني وصعوبة الوصول لمعلوماتك المطلوبة عبر موقع الجامعة الالكتروني جعلك تلجأ لإستخدام هذا النظام” “The difficulty of contacting the university by phone or email, and accessing your needed information on the university website were the reasons to use ArabChat”.
8. “لقد ساهم النظام في توفير جهدك و وقتك .” “The agent saves you time and effort”.
9. “ لايوجد خدمة مثيلة باللغة العربية لأي جامعة ,كلية, أو لشركة و أيضا لايوجد نظام اسئلة وأجوبة باللغة العربية” “There is no Arabic university, college or company offering the same services, even there is no question answering system in Arabic”.
10. “كان النظام يشجعك بالاستمرارعلى الحديث” “The agent encourages you to carry on with the conversation”.
11. “تقييمك الإجمالي للنظام بأنه ممتازا” “Your overall rating for this service is excellent”.
12. “سوف تنصح أصقائك باستخدام هذا النظام” “You will recommend your friends to use the ArabChat system”.
13. “ أنت تفضل استخدام هذا النظام بدلا عن التحدث مع الشخص المسؤول في الجامعة” “You prefer to use ArabChat rather than speak with a human advisor”.
14. “سوف تعيد إستخدام النظام في المستقبل” “You will re-use this service in the future”.

The Enhanced ArabChat online questionnaire system will not accept the submission of any questionnaire without completing all the questions.

The subjective evaluation results

159 of 203 of the Enhanced ArabChat experiment 1 users submitted the online questionnaire. Table 8 presents the Enhanced ArabChat online questionnaire results.

TABLE VIII. THE ENHANCED ARABCHAT ONLINE QUESTIONNAIRE RESULTS

#	“Agree” distribution (Percent)	“Neutral” distribution (Percent)	“Disagree” distribution (Percent)
1	142 (89.3%)	11 (6.9%)	6 (3.8%)
2	140 (88.1%)	19 (11.9%)	0 (0%)
3	122 (76.8%)	29 (18.2%)	8 (5%)
4	153 (96.2%)	6 (3.8%)	0 (0%)
5	96 (60.4%)	44 (27.7%)	19 (11.9%)
6	51 (32.1%)	56 (35.2%)	52 (32.7%)
7	126 (79.2%)	33 (20.8%)	0 (0%)
8	115 (72.3%)	37 (23.3%)	7 (4.4%)
9	152 (95.6%)	7 (4.4%)	0 (0%)
10	48 (30.2%)	46 (28.9%)	65 (40.9%)
11	107 (67.3%)	34 (21.4%)	18 (11.3%)
12	95 (59.7%)	45 (28.3%)	19 (11.9%)
13	103 (64.8%)	29 (18.2%)	27 (17%)
14	109 (68.6%)	34 (21.4%)	16 (10.1%)

Discussion

According to Table 8, the questionnaire questions will now be discussed based upon the evaluation aims as discussed before:

- The Enhanced ArabChat user interface evaluation: the user interface was evaluated using item number 1. 89.3% of users agreed that the Enhanced ArabChat user interface was suitable.

- The Enhanced ArabChat usability evaluation: the Enhanced ArabChat usability was evaluated through 3 items in the questionnaire which are 4, 7, and 8. 96.2% of users agreed that they experienced no technical problems while using the Enhanced ArabChat. 79.2% of users agreed that difficulty contacting the university by phone or email, as well as difficulty accessing their needed information on the university website were the reasons that caused them to use the Enhanced ArabChat. Finally, 72.3% of users agreed that the agent saved them time and effort.

- The Enhanced ArabChat naturalness evaluation: the Enhanced ArabChat’s naturalness has been evaluated through 3 items: 3, 6, and 10. 76.8% of users agreed that the Enhanced ArabChat’s responses were clear and understandable. Only 32.1% of users mentioned that the Enhanced ArabChat’s interaction was realistic and believable. 40.9% of users disagreed with the notion that Enhanced ArabChat encouraged them to carry on with their conversation. This inability to encourage further conversations might be due to the response scripting, which fails to encourage users to continue conversations after firing certain rules. This might provide evidence that the large number of question-based utterances in experiment 1 was due Enhanced ArabChat responses not encouraging users to keep conversations going.

- The applied domain coverage evaluation: the applied domain coverage has been evaluated through item number 2. 88.1% of users agreed that Enhanced ArabChat was able to provide all of their requested information, indicating that the applied domain coverage topics were good enough to cover ASU students’ issues.

- The Enhanced ArabChat interaction speed evaluation: the interaction speed of Enhanced ArabChat has been evaluated through item number 5. 60% of users agreed that the elapsed time taken by Enhanced ArabChat to handle their utterances was reasonable.

- The availability of Similar Arabic agent evaluation: The availability of similar Arabic CAs was evaluated through item number 9. 95.6% of users agreed that there is no Arabic university, college or company offering the same services. This high percentage carries two meanings behind it. First, Enhanced ArabChat might be considered the first Conversational Agent responsible for handling user utterances in the Arabic language. Second, it might reflect the users’ inability to differentiate between CAs and QA(Question Answering) systems. According to Table 5, 74.292% of user utterances were questions. As a result, they might consider it as a QA system due to the lack of experience using similar systems. This fact supports the third factor emphasizing the large number of question-based utterances are due to the users' confusions about whether the Enhanced ArabChat is a QA or a CA.

- The user general satisfaction evaluation: the general satisfaction of the Enhanced ArabChat users was evaluated through item numbers 11, 12, 13, and 14. 67.3% of users agreed that their overall rating for Enhanced ArabChat was excellent, while 59.7% agreed to recommend Enhanced ArabChat to their friends. 64.8% of users prefer to use Enhanced ArabChat rather than speak to a human advisor. Finally, 68.6% of users confirmed they would use the Enhanced ArabChat for future needs.

B. Experiment 2

As discussed in experiment 1, 74.292% of the utterances were question-based. This high percentage of question-based utterances might be caused by different factors as discussed in experiment 1. All of the mentioned factors were investigated in experiment 1 apart from the factor that the Enhanced ArabChat scripting engine does not deal well with non question-based utterances. Therefore, this experiment has been conducted on Enhanced ArabChat in order to investigate this factor. Given this, the evaluation of Enhanced ArabChat based on experiment 2's results will be limited to the metrics that related to the engine only.

Experiment 2 methodology

In this experiment, 17 users were asked to have a conversation with Enhanced ArabChat. The 17 users were randomly selected to converse with Enhanced ArabChat. All users were students in ASU from different courses. This experiment focused only on the utterance type (question or non-question) and its effect on continuing conversations. The users were requested to chat with the Enhanced ArabChat by entering non-question-based utterances as much as possible. In other words, they were required to avoid asking questions. The number of utterances that a user should enter was not determined. Therefore, different users entered a different number of utterances.

Experiment 2 results

In this experiment, the Enhanced ArabChat handled 104 utterances from 17 users. The number of classified utterances as questions and non question are presented in Table 9. The results of experiment 2 will be discussed with the evaluation of Enhanced ArabChat based upon these results in the next section.

TABLE IX. QUESTION VS. NON-QUESTION UTTERANCES BY ENHANCED ARABCHAT

Utterance classified Type	**Number of classified utterances (percent)**
Question-based	18 (17.3076 %)
Non Question-based	V. (82.692 %)

The Enhanced ArabChat evaluation based on experiment 2 results

The evaluation of the Enhanced ArabChat for this experiment will deal only with metrics that meet the discussed factor (the engine factor). Therefore, only the objective evaluation will be conducted including the "Glass box approach" evaluation. The "Glass box approach" will only be used to evaluate the Enhanced ArabChat "Utterance Classification" module and the scripting engine.

1) The objective (Glass box) approach evaluation

a) Utterance classification evaluation

Evaluation aim

This evaluation aims to evaluate the performance of the "Utterance Classification" module.

Evaluation methodology

The Enhanced ArabChat handled 104 utterances from 17 users in experiment 2. These utterances were classified into question-based and non-question-based utterances, as presented in Table 9. A manual classification process for the 104 utterances was conducted in order to evaluate the real correct module performance.

Evaluation results

The real number of question-based and non question-based utterances is presented in Table 10.

TABLE X. THE REAL (QUESTION VS. NON-QUESTION) UTTERANCES (MANUAL CHECKING)

Utterance classified Type	**Number of classified utterances (percent)**
Question-based	14 (13.4615 %)
Non Question-based	90 (86.5384 %)

Discussion

Table 9 presents the classified results of the Enhanced ArabChat. According to the table, 82.6923% of the total utterances are non-question-based. However, Table 10 presented the correct number of non-question-based utterances as 86.5384% of total utterances (manual checking). As a result, the Enhanced ArabChat classifier can be considered acceptably accurate for the two types of utterances (question and non question).

b) The Enhanced ArabChat scripting engine

Evaluation aim

The evaluation aim is to determine the RMUT of the Enhanced ArabChat.

Evaluation methodology

To conduct this evaluation, the RMUT equation has been used.

Experiment Results

The results show that the average of RMUT for the 17 users is 72.12%.

Discussion

The results reported in the previous section show that 72.12% of the Enhanced ArabChat users' utterances were matched. This technique cannot reveal if the matching led to a successful conversation or a failed conversation. However, the RMUT, as discussed earlier, gives a general overview of the Enhanced ArabChat scripting engine's performance.

VI. Conclusion

This paper described the Enhanced ArabChat which it is a complement of the first version of ArabChat [4]. Therefore, all the developed features in the first version are also included in the Enhanced ArabChat. In addition, some new features that have been revealed from evaluating the first version of ArabChat are added to improve the agent performance. These new features were "Utterance Classification" and "Hybrid Rule" as described in this paper. Integrating these new features ("Utterance Classification" and "Hybrid Rule") has changed the engine working methodology of the Enhanced ArabChat as discussed in this paper. These changes might reflect positively the performance of the Enhanced ArabChat.

A comprehensive evaluation methodology consisting of objective and subjective approaches has been used to evaluate the Enhanced ArabChat. The objective approach has been conducted through automatic evaluation techniques and manual analysing. The "Glass box" approach evaluated the Enhanced ArabChat components individually. The Enhanced ArabChat obtained a 67.836% of RMUT. This result can give a general overview of Enhanced ArabChat performance, but it does not give a full indicator of its performance. Hence, a new comprehensive evaluation technique for CAs should be modelled and developed. The subjective evaluation showed that 67.3% of users who submitted the questionnaire agreed that their overall rating for Enhanced ArabChat was excellent, and 64.8% of them prefer to use it rather than speak with a human advisor.

It has been observed in experiment 1 that users entered more question-based utterances than non-question-based ones. This might be due to four factors, including the nature of the selected domain, the engine, the user and the scripts. In experiment 1, three of these factors have been discussed and verified as accurate reasons for this problem: the nature of the scripted domain, the user and the scripts (the Enhanced ArabChat responses). In experiment 1, it was noticed that non-serious users negatively affected the calculated user satisfaction by chatting with the Enhanced ArabChat in an indecent manner (not covered by the "Bad words" context). Also, these non-serious users tried entering many questions just to trick the Enhanced ArabChat. Moreover, as discussed in the subjective evaluation, 95.6% of Enhanced ArabChat users agreed that this was the first time they used such a service (ArabChat information point advisor). Thus, the numerous question-based utterances might be to the fact that users cannot differentiate between CAs and QA systems.

Therefore, experiment 2 was conducted for the Enhanced ArabChat to check the fourth factor (the engine factor) that might have caused the large number of question-based utterances. Experiment 2 confirmed that Enhanced ArabChat successfully dealt with non-question-based utterances, as the reported user satisfaction rate was 70.488%. Consequently, it was concluded that Enhanced ArabChat scripting engine can deal with non-question-based utterances. This evidence led to the rejection of the fourth potential factor.

Generally, chatting with a CA does not mean that a user will keep entering either questions or non-questions only. The natural conversations between a user and a CA should consist of both (questions and non-questions). Nevertheless, the amount of question and non-question utterances might be based on the following factors:

1) The topical nature of a CA's applied domain; for instance, an entertainment domain might differ from an information point advisor.

2) The users, if they are familiar with the nature of a CA. It can be concluded from experiment 1 of Enhanced ArabChat that many users consider it a question answering system. As a result, a lot of questions were entered. Also, 92.3% of experiment 1's users confirmed that they had never used any similar service before, which points to a lack of experience in handling these services.

3) The way a CA forms its response might also encourage a user to ask questions or continue chatting with non-question utterances.

According to the two conducted experiments (experiment 1 and experiment 2) and the evaluation of the Enhanced ArabChat based on these experiments' results, it can be concluded that the Enhanced ArabChat successfully handled conversations for ASU students.

Acknowledgment

The authors are grateful to the Applied Science Private University, Amman, Jordan, for the full financial support granted to this research.

References

[1] Turing, A., Computing machinery and intelligence. Mind, 1950: p. pp 433-60.

[2] Turing, A., Computing machinery and intelligence. MIT Press, 1995: p. 11-35.

[3] O'Shea, K., Z. Bandar, and K. Crockett, A Novel Approach for Constructing Conversational Agents using Sentence Similarity Measures. 2008.

[4] Hijjawi, M., et al. ArabChat: An Arabic Conversational Agent. in proceeding of the 6th International Conference on Computer Science and Information Technology (CSIT). 2014. Amman, Jordan: IEEE Explore.

[5] Crystal, D., Dictionary of linguistics and phonetics., Blackwell., Editor. 2008.

[6] Habash, N., Introduction to Arabic Natural Language Processing, ed. U.o.T. Graeme Hirst. 2010: Morgan & Claypool.

[7] Sammut, C. and D. Michie, InfochatTM Scripter's Manual, Convagent Ltd. . 2001: Manchester.

[8] Weizenbaum, J., ELIZA-A computer program for the study of natural language communication between man and machine. Communications of the ACM., 1966. Vol 10.: p. PP 36-45.

[9] Wallace, R. ALICE: Artificial Intelligence Foundation Inc. . 2008 [cited; Available from: http://www.alicebot.org.

[10] Timothy, B. and G. Toni, Health dialog systems for patients and consumers. J. of Biomedical Informatics, 2006. 39(5): p. 556-571.

[11] Maragoudakisa, M., et al., Natural Language in Dialogue Systems, a case study on a medical application. , in Proceedings of Panhellenic

Conference with International Participation in Human–Computer Interaction. 2001: Greece. . p. 197–201.

[12] Shaalan, K., Rule-based Approach in Arabic Natural Language Processing. 2010.

[13] Sammut, C., Managing Context in a Conversational Agent. Electronic Transactions on Artificial Intelligence, 2001.

[14] Ong Sing, G. and F. Chung Che, The design of interactive conversation agents. WSEAS Trans. Info. Sci. and App., 2008. 5(6): p. 901-912.

[15] Hijjawi, M., Z. Bandar, and K. Crockett. User's utterance classification using machine learning for Arabic Conversational Agents. in proceeding of the 5th International Conference on Computer Science and Information Technology (CSIT). 2013: IEEE Explore.

[16] Hijjawi, M., Z. Bandar, and K. Crockett, A Novel Hybrid Rule Mechanism for the Arabic Conversational Agent ArabChat. Global Journal on Technology, 2015(Issue 8): p. 185-194.

[17] ASU. Applied Science University. 2011 [cited; Available from: www.asu.edu.jo.