Building an Arabic Application employing Information Extraction Technology

Daoud Maher Daoud
Alzaytoonah University of Jordan
daoud_m@alzaytoonah.edu.jo
&
GETA, CLIPS, IMAG; daoud.daoud@imag.fr

Abstract

In this paper, we describe a SMS-based information system called CATS, which allows posting and searching through free Arabic text using Information Extraction (IE) technology. We discuss the challenges of applying IE technology for unedited real Arabic text. In addition, we describe the structure of this system and our approach to produce an open robust system capable of including more sub domains with the minimum effort.

Keywords: Information Extraction, Arabic Language Processing, Classified Ads, Attribute Based Searching.

1 Introduction

Natural language is considered the simplest technique of human-machine interaction. It is suitable for naïve users who know the task domain well. However, building a robust commercial application that employ natural language requires restricted domain where we have control over linguistic and world knowledge.

Information Extraction (IE) is a comparatively new technology within the more general field of Natural Language Processing. IE is the process of identifying relevant information where the criteria for relevance are predefined by the user in the form of a template that is to be filled [9]. The current development in the field of IE is owed to the Message Understanding Conferences (MUCs). In this competition English has always been the unique target language, with the exception of MUC-6 (MET-1) where Spanish and Chinese were considered as well [10]. IE systems are usually designed for a specific domain, and the types of facts to be extracted are defined in advance [11]. Most of the researchers believe that the IE technology is promising and pertinent to a wide range of fields, despite that much of the research have been directed toward news items found in the web. IE systems are a key factor in encouraging NLP researchers to move from small-scale systems and artificial data to large-scale systems operating on human [7].

In this paper, we will describe our efforts for employing IE technology in a SMS based information system called CATS, which uses Arabic as an interaction language for connecting sellers and buyers through SMS in the classified ads domain.

2 Background

The Classified Ads through SMS (CATS) system is a SMS based classified selling and buying platform. Users can send classified ads of the articles/goods they would like to sell, and can search for the goods/articles they desire in the platform economically and while moving. It provides the user with a natural language interface where the user can specify his/her request by sending SMS text in Arabic to an assigned short number.

SMS, **or Short Message Service,** is becoming the most popular channel for exchanging information. The most important factor that explains this enormous success is its simple, immediate, and confidential way to communicate. Moreover, it has played a major role in narrowing down the digital gap caused by low level of internet penetration in some countries. As an example, SMS enables communication with more than 1.5 million Jordanian subscribers anywhere, anytime, and hence offers unmatched service coverage, beyond even that of the Internet as mobile phone penetration is much higher than Internet usage.

In the same context, classified ads are an effective way of connecting buyers and sellers. Normally, they are concise with a limited but specialized vocabulary. They are rich in proper nouns, nouns, hint words, and numerical values.

The CATS system can handle both unstructured free Arabic SMS texts and structured data stored in a relational database. When the CATS system receives

a text it extracts the relevant information and distinguish between the “posting text” and “search text.” Both of them are processed similarly by filling previously designed templates. For a “posting text,” the template is stored in a database, and for a “search text,” the template is used to build a query to retrieve information which resides in the database.
The current version of the CATS system is in Arabic and is restricted to classified ads domain. The cars and real estate sub-domain are implemented in this version. However, the system is structured to adapt other sub-domains. Moreover, we have plans to produce a multilingual version of the system.

3 Information Extraction and Arabic

Not all languages have received equal investment in linguistic resources and tool development [11]. As an example, most of the research published on IE discussed problems related to English, which is a resource-rich language. In the same context, some of the existing English based IE systems performance is comparable to human experts. On the other hand, Natural Language Processing (NLP) in the Arabic language is still in its initial stage compared to the work in the English language [12].
Regarding Information Extraction, Arabic was not one of the languages considered in the MUCs events. However, Arabic is supported, along with English and Chinese, in the Automatic Content extraction (ACE) program that is operating under the DARPA Program in Translingual Information Detection, Extraction, and Summarization (TIDES). The ACE research objectives are viewed as the detection and characterization of Entities, Relations, and Events [13]. Annotation of named entity is the core of this project. Full-scale Chinese annotation is well underway, while Arabic annotation is just beginning [13].
The common practice of automatically extracting information has been through using of templates, which specify what information should be harnessed. Accordingly, for example, a template for a car classified ads scenario might specify fields such as "Car Make", "Car Model", "Year", "Color", "Mileage", “Price”, “Phone”. The IE engine would then try to fill these fields similar to filling the information in a database. This task (referred to as the Template Element task) has been examined in detail in MUCs, aiming at accurately identifying names, dates, and organizations in a text.
Despite these developments, building a large-scale information system based on IE that supports Arabic poses new challenges that do no exist in English or other resource-rich languages.

- Classified ads are rich in proper names, which in Arabic are not distinguished by using upper case letters like English. This makes it not nearly as easy to locate them in Arabic text as in English text [16].

- The Variations of spelling of the Arabic text caused by its complex orthography adds more challenges to the processing of the Arabic text. As an example people tend to interchange between the Alef “ا” , “أ” and “إ” in their writing, also between the Ha’ “ه” and Ta’ “ة’”, and between Ya’ “ي” and Alef-Maqsoura “ى”. In Arabic, spaces are normally used to separate words. Most of Arabic letters are connected from both sides (cursive writing system), causing them to have different shapes depending on their positions (first, middle, or last). But some letters “و”, “ر” ,”ز” , ”د”, ”ى” and ”ذ” can be connected only from the right side making their shapes unchanging at any position of the word. After any of these letters, people tend to insert a space or simply drop it (e.g., “أبو بكر” or “أبوبكر” {Abu-Baker}).

- The inconsistency of the Arabic spelling of transliterated proper nouns is a major challenge. This appears frequently in the classified ads text where many of the proper names (car make and model as an example) are transliterated from other languages. This phenomenon is noticeable within unedited and spontaneous classified ads, reflecting the cultural and educational background of the text writer. As an example the car-make CITROEN could have different spelling in Arabic:{SATARWEN}“سترون”,{SA:TERWEN} “ساترون “ , {SATERWE:N} “ستروين”, {SA:TERWE:N} ”ساتروين ”, {SE:TERWE:N} “سيتروين”.

- Arabic Language uses a diverse system of prefixes, suffixes, and pronouns that are attached to the words, creating composite forms that further complicate text manipulation. For instance, articles such as "an" and "the" are not separate words as they are in languages like English but are actually appended to the words to which they refer (for example, "their two cars" is written as a single token, سيارتيهم).This can cause ambiguity of forms (especially if short vowels are omitted). As an example, in Arabic,”سترون ” has two interpretations, which are (you will see) and (CITROEN a name of a car brand).

Therefore, Arabic's rich morphology and complex orthography present unique challenges for analysis, which requires significant pre-processing before it can be accurately indexed, searched, or put through any other text manipulation.

4 The CATS System

4.1 Overview

The CATS system is an information system that uses IE technology. The goal of this application is to enable SMS users to post or search for classified ads in Arabic. It has two main functionalities: the submission for selling items and the answering of users' queries through natural language interaction. The system receives an entry in full text without a pre-specified layout, recognizes the various relevant entries, and produces a logical representation for further processing. We have two types of users' requests:

- Posting type in which the user is a potential seller.
- Searching type in which the user is a potential buyer.

Suppose a user wants to sell his car, he can simply type a SMS message and send it to a specified short number (Figure 1). Similarly, the user can express his search request and send it to the same number. As show in Figure 2 the system is capable of performing exact (e.g., Japanese) and specified ranges (e.g., the price should be less than 2500 JD) attribute-based search. If the system finds records that match his request, the user will get a list of cars with contacts numbers (Figure 2). Otherwise, the user will get a notifying message telling him that no match found at this time with the possibility to get results later on.

In addition, The CATS system implements a leveled based strategy for searching. As an example, consider this query "I am looking for Clio 1999." Suppose the system fails to find any match, it will look for all cars manufactured by Renault cars maker. In the same way, if it fails to find any match it will look for any French car.

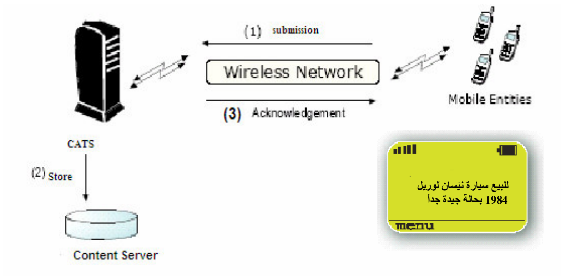


Figure 1 Sending a selling classified ad "For sale Nissan Laurel 1984 in good condition"

.

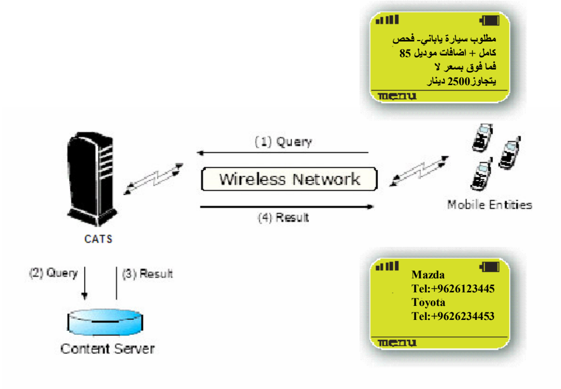


Figure 2 Making a query and getting results through CATS system. "A Japanese car is wanted, full check up, full option, above 1985; the price should be less than 2500 JD."

Finally, the current version of the CATS system includes the cars and real estate sub-domains. However, we took into consideration in the design to make the system customizable to include other sub-domains with the minimum possible effort.

4.2 Template Design

Developing a high-quality system requires a systematic approach. We started the development by collecting a corpus for each particular sub-domain. This corpus was collected from the web sites that provide unedited Arabic classified ads services. By having access to this corpus, we have been able to study the used patterns and even to anticipate patterns that were not seen in the corpus. More to the point, the corpus enabled us to depict the lexicon, styles and types of queries that interest users. We also made decisions on what is relevant and what is not to a particular domain. Then we began putting our template, which reflects our conceptual view of the relevant knowledge embedded in the free text of the classified ads.

We have adopted the object-attributes to model the templates. This representation acts as medium between free text and structured database. They abstract our conceptual view for a particular domain.

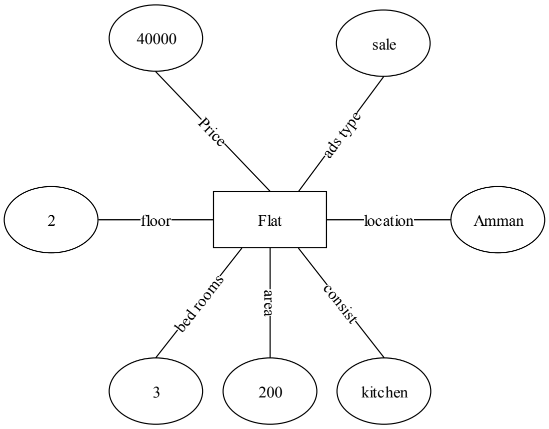


Figure 3 An example of object-attributes model

For example a flat has a location, consists of parts, has an area, has a number of bedrooms, has a floor number, has a price and either it is "for sale" or it is "wanted."

Figure 3 shows that there is a main object, which has a value "flat" along with its attributes.

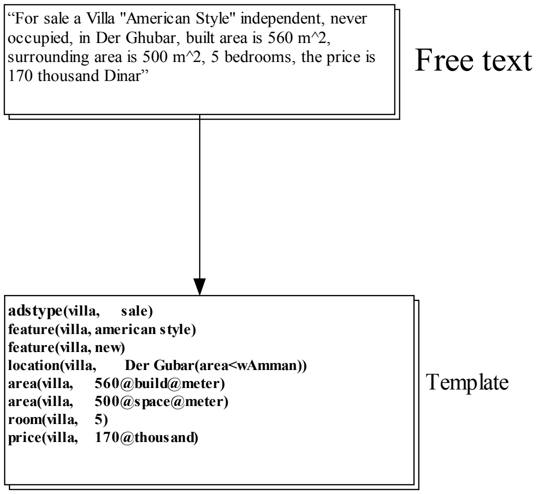
In real estate sub-domain, the main object is "flat", "land", "shop", "building", etc.

We also defined a set of specifieres for each sub-domain to give more details or to put some restriction on the attributes values. Specifically, they are used in normalization of numerical attributes values and in capturing mentioned ranges used to perform attribute based searching. As an example, in the following expression:

Price (car, 5000@less)

"@less" means that the price of the car should be less than 5000.

In figure 4, "@meter" means that the unit of measuring is meter. Similarly, the "@build" used to indicate building area and "@space" indicates surrounding area. When a numerical value is attached to "@thousand", this number should be multiplied by 1000 for further processing.



"For sale a Villa "American Style" independent, never occupied, in Der Ghubar, built area is 560 m^2, surrounding area is 500 m^2, 5 bedrooms, the price is 170 thousand Dinar"

adstype(villa, sale)
feature(villa, american style)
feature(villa, new)
location(villa, Der Gubar(area<wAmman))
area(villa, 560@build@meter)
area(villa, 500@space@meter)
room(villa, 5)
price(villa, 170@thousand)

Figure 4 an example of a template of a classified Ad

4.3 The Dictionary

The dictionary is the backbone of the CATS system. It maps the domain specific Arabic words to normalized forms. This dictionary also encode morphological, syntactic, and semantic features for each word used for parsing of the classified ads text, in addition to the ontological information which are necessary to perform leveled based searching. As shown in figure 5, the different Arabic words for the car make CITROEN are mapped to one normalized form, which is the only form to be used in filling the slots of the templates. This structure will minimize the effect of the alternative representations of text (including different orthographic forms, spelling errors, and abbreviations) on the overall performance of the system, specifically in the searching process. Dictionary coverage is an especially challenging problem, since classified ads can be filled with all manner of jargon, abbreviations, and proper names, not to mention typos and fragmented phrases instead of fully-formed proper Arabic sentences. Currently, the dictionary contains 30,000 entries. Most of them are generated automatically by applying certain normalization algorithms.

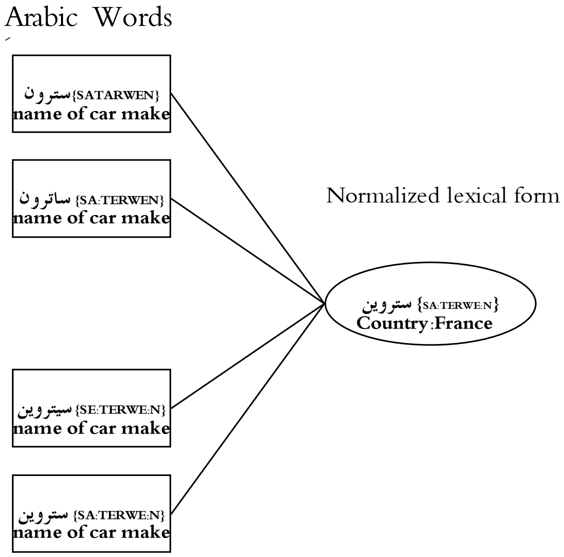


Figure 5 an entry of the Dictionary

4.4 Natural Language Processing for the Classified Ads

When dealing with classified ads text, it is important to recognize deferent entities such as car makes (e.g., “Honda”), car models (e.g., “Accord”), locations (e.g.,” Amman”), property types(e.g., “flat”), vehicle types (e.g., “saloon”) . It also includes the identification of other patterns such as numerical expressions: e.g., price, area, motor size, etc. To extract text attributes (location names, cars makers names, etc.) we use mainly the lexical lookup approach, and for numerical attributes, we use mainly the rule-based approach

We can view a classified ad as a group of information entities assembled in a natural language structure. A systematic approach is necessary for the analysis of the free text. We believe that the development of a methodology to specify the language grammar used in a particular domain is necessary for building efficient Information Extractions systems. IE needs interface specification between NL and domain knowledge [17].

As we have shown above, the first step was the template design. It is a link between the unstructured free text and the structured database. By defining the sets of valid objects, relations, properties and attributes within the template, we set the boundary for each particular domain. The template design reflects the corpus or what concern people in a specific domain. It is also a natural language independent representation of the information.

The analysis of a received classified ad is affected by the following:

- The template design
- The structure of the natural language.

The core problem of building IE systems is to identify general mapping between text fragments and template descriptions [17]. Additionally, an IE system crucially needs to capture various ways in which this information may be described.

4.4.1 Knowledge Component

The main function of this methodology is to specify diagrammatically the natural language patterns using knowledge components (KC), which are entities that embed morphological, syntactic, and semantic information that can be identified in the natural language text from its unique function in a particular domain.

Through the Knowledge Component, we can specify adjacency relations, priority levels, constraints, and conditions.

A car classified ad might contain the following knowledge components:

- Main object
- Ads_type
- Price
- Feature
- Motor_size
- Year
- Country
- Color
- Model
- make

Each component has one or more surface structure (variants), but serves one knowledge function. As an example, each of the following phrases:

محرك 1500 سي سي	Motor is 1500 cc
1500 سي سي	1500 cc
سعة المحرك cc1500	motor size is 1500
1.5 ليتر	1.5 L

has a different syntactic structure, but all hold the same knowledge, which is described by:
Motor_size (car, 1500)

In the same context, consider the following phrases found in the real estate sub-domain that is describing a property:

تصلح لبناء بيت	……fit to build a house,………
تقع وسط فلل	……..Surrounded by villas ……….

سكني	………residential ……..
في منطقة سكنية	………Located in a residential area,………
لبناء بيت او فيلا	….. to build a house or villa ………..

All of the above phrases indicate that the property is residential (not commercial or industrial) which is expressed in the template as:

Type (land, residential)

Defining a Knowledge Component involves specifying the knowledge function and the surface structure(s) that might appear in a free text within a particular domain.

Each surface structure consists of simple components derived from the dictionary and/or composite ones.

As shown in Figure 6, the primary constituent of motor size KC is the number. It shows that the number has non-compulsory right adjacency relations to a right quantifier and a motor size hint components. To the left, it has also non-compulsory relations with a left quantifiers and a motor size unit word. The quantifiers and the motor size hint are composite components that have to be defined. On the other hand, the motor size unit constituent is atomic and it is derived directly from the dictionary.

The constituents of any component are categorized into hint constituent, and control constituents. Hint constituent help in identifying the primary constituent, while control constituent has an influence in the primary constituent. As an example to recognize that a number in a car classified ad is a motor size, it should preceded by a motor size hint component such as “the size of the engine” or followed by the appropriate unit such as “cc”. On the other hand, the negation constituent in the “right quantifier” changes the primary one from “less” to “more” and vise versa, which is leading to changing the knowledge function of the component.

The knowledge components have a hierarchical structure (figure 7), where the classified ad itself is considered the top component consisting only of other knowledge components, where no atomic constituent such as dictionary component are allowed.

Likewise, in the top knowledge component we have to define the primary component. As an example, in a car classified ad the primary knowledge component is the “vehicle,” while in the real estate sub-domain it is the “property.”

Likewise, we defined the knowledge components for each particular domain and all their constituents.

These KCs encapsulate domain specific knowledge along with natural language structures.

They show our parsing strategy in focusing to extract domain dependent information from the free text regardless of the surface structure they appear in.

Although some of the components used for modeling the system are domain independent, they still serve the main function of a particular domain.

This methodology defines the partial parsing for a certain domain by indicating domain relevant entities, relations, attributes, and different surface structures. Therefore, it is steering the parsing process, and by describing these KCs with finite-state grammars, which are processed easily, robustly and rapidly.

This systematic methodology of development of an IE system such as CATS proved to be feasible. Consequently, we were able to build up a robust solution in a shorter time. This methodology is also very useful in expanding the system to accommodate more sub- domains.

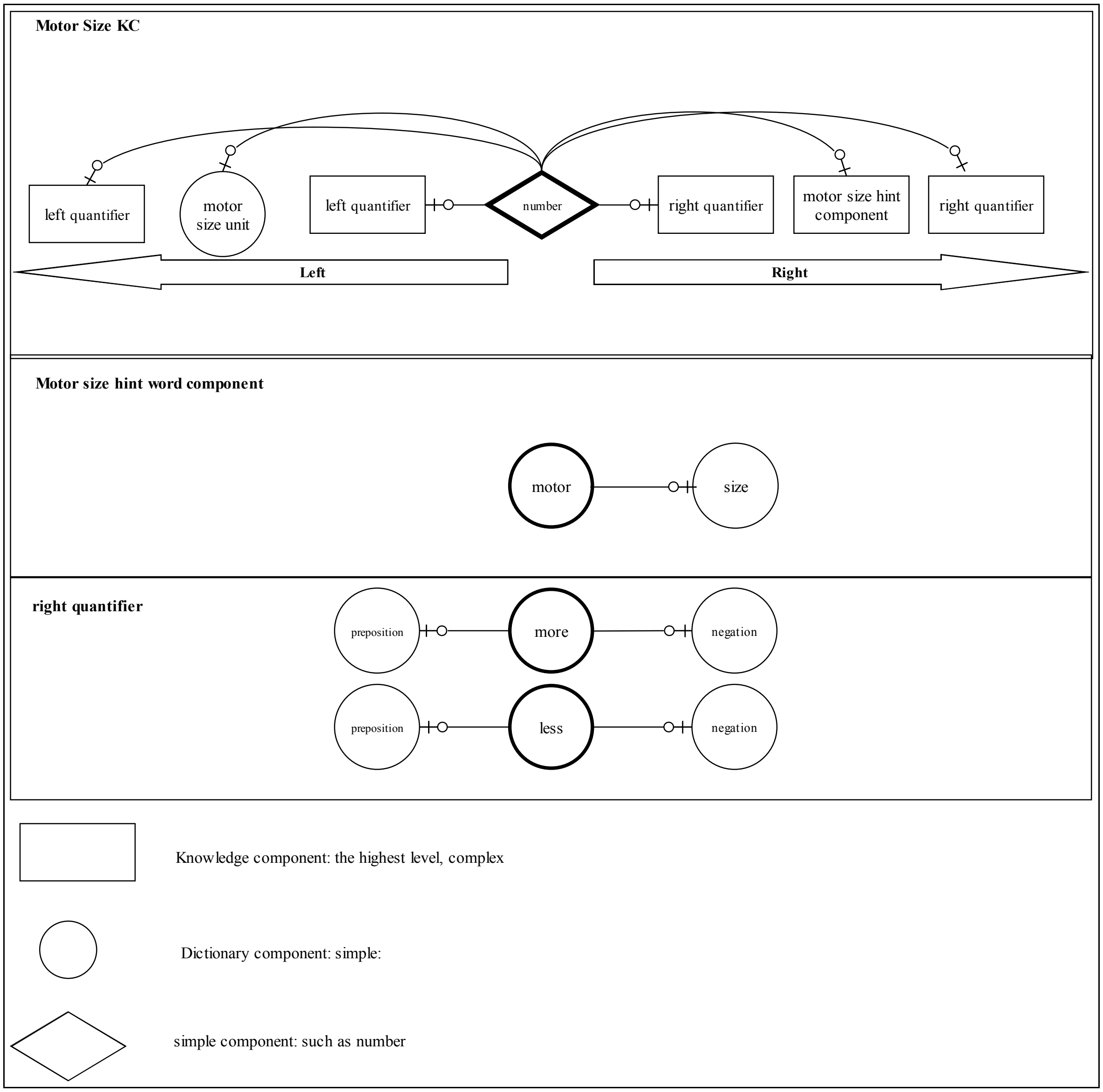


Figure 6 Examples of components definition

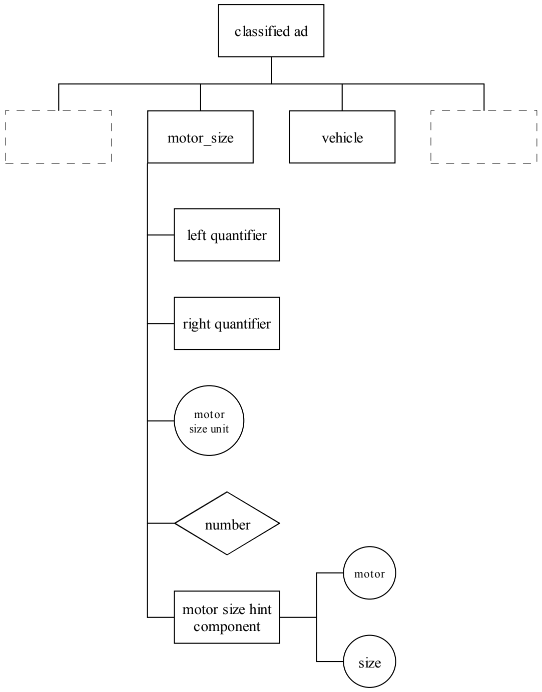


Figure 7 The hierarchical structure of the knowledge components

5 Evaluating the CATS system

Because the CATS system is targeting end users, we preferred to evaluate the system by surveying them directly. Initially, the system has been deployed with the cooperation of the largest mobile operator in Jordan (FASTLINK) to perform this evaluation. We have explained the system to a sample of around 200 users from different backgrounds, asking them to test the system by sending posting (for sale) and searching (looking to buy) SMS messages.

Generally, the feedback was positive: 95% of the participants said that result were accurate. The rest said that the results should be more precise. We have noticed that 70% of the messages are of the search type. We also noticed that the quality of the sent text is good with little spelling errors.

These results encouraged FASTLINK to deploy commercially the CATS system as part of their premium services.

6 Future Work

Our next work will focus on using the methodology described above for generating the domain specific rules automatically from the knowledge components definitions in order to include more sub-domains easily and efficiently.

We will also work on making the CATS system Multilingual by enabling people to submit and search classified ads in any language using Universal Networking Language framework (UNL).

7 Conclusions

The purpose of this paper has been to report a commercial Arabic mobile application employing Information Extraction technology. We have shown the functionalities and architecture of the CATS system.

We have also shown the importance of having a systematic approach of domain parsing. Therefore, we have introduced the Knowledge Components (KC) as methodology of engineering the parsing process.

.

References

[1] B. Vauquois & S.Chappuy (1985) Static Grammars. Proc. Conf. On theoretical & methodologies issues in MT (TMI-1), Colgate Univ., Hamilton, N.Y.,August 1985.

[2] Daoud Maher (2005) Arabic Deconversion in the framework of the Universal Networking Language, Cicling 2005, Mexico, Feb 2005.

[3] Daoud Maher (1999) Arabic Deconversion: Problems and Prospects, UNL 99 European workshop, Perugia, Italy, July 1999.

[4] UNU/IAS (1999) Deconverter Specifications. UNU/IAS UNL Center, Http://www.undl.org

[5] UNL center UNDL Foundation (2003). The Universal Networking Language specification. Http://www.undl.org.

[6] Constructing Better Document Vectors Universal Networking Language (UNL) (2002),Chirag Shah, Bhoopesh Chowdhary, Pushpak Bhattacharyya, Proceedings of International Conference on Knowledge-Based Computer Systems (KBCS) 2002.

[7] Jim Cowie and W. Lehnert. Information extraction. *Special NLP Issue of the Communications of the ACM*, 39(1):80-91, January 1996.

[8] Appelt, D. E., Israel, D. J., Introduction to Information Extraction Technology. A Tutorial prepared for the 16th International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, Sweden, August 1999.

[9]Grishman, R. et al, multilingual Information Management: Current Levels and Future Abilities, chapter 3: Cross-lingual Information Extraction and

Automated text Summarization. New York, April 1999. Also available at http://www-2.cs.cmu.edu/~ref/mlim/

[10] Dini, L. Parallel Information Extraction Systems For Multilingual Information Access, In Proceedings of the Third European Robotics, Intelligent Systems & Control Conference (Euriscon 98), June, 1998, session 6 (Adaptive and Multilingual Information Extraction Systems). http://citeseer.ist.psu.edu/dini98parallel.htm

[11] Ellen Riloff, Charles Schafer, and David Yarowsky. ``Inducing Information Extraction Systems for New Languages via Cross-Language Projection." In Proceedings of COLING 2002.

[12] Bassam Hammo, Hani Abu-Salem, Steve Lytinen, Martha Evens. QARAB: A Question Answering System to Support the Arabic Language. Computational Approaches to Semitic Languages workshop, 11th July 2002, University of Pennsylvania.

[13] Stephanie Strassel; Alexis Mitchel. Multilingual Resources for Entity Extraction, Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition.

[14]J.R. Hobbs, The generic information extraction system. In Proceedings of the fith Message Understanding Conference, (MUC5), Morgan Kaufman, 1993.

[15] Y. Wilks and R. Catizone, "Can we make information extraction more adaptive", in M.T.Pazienza (ed.), "Information Extraction", Springer, 2000.

[16] Abuleil, S., and Evens, M., 1998. “Discovering Lexical Information by Tagging Arabic Newspaper Text”, Workshop on Semantic Language Processing. COLING-ACL ’98, University of Montreal, Montreal, PQ, Canada, Aug. 16 1998, pp. 1-7.

[17] Guenter Neumann and Feiyu Xu, Intelligent information extraction, 16th European Summer School in Logic, Language and Information, Université Henri Poincaré Nancy, France9-20 August, 2004