Extracting Arabic Composite Names Using Genitive Principles of Arabic Grammar

HUSSEIN KHALIL, TAHA OSMAN
Nottingham Trent University
School of Science & Technology
{hussein.khalil2007, taha.osman}@ntu.ac.uk

MOHAMMED MILTAN
Misurata University
Arabic Department, Faculty of Arts
{mmlitan@gmail.com}

ABSTRACT

Named Entity Recognition (NER) is a basic prerequisite of using Natural Language Processing (NLP) for information retrieval. Arabic NER is especially challenging as the language is morphologically rich and has short vowels with no capitalisation convention. This paper presents a novel rule-based approach that uses linguistic grammar-based techniques to extract Arabic composite names from Arabic text. Our approach uniquely exploits the genitive Arabic grammar rules; in particular, the rules regarding the identification of definite nouns (معرفة) and indefinite nouns (نكرة) to support the process of extracting composite names. Based on domain knowledge and Arabic Genitive Rules (AGR), the developed approach formalises a set of syntactical rules and linguistic patterns that initially use genitive patterns to classify definiteness within phrases and then extracts proper composite names from the unstructured text. The developed novel approach does not place any constraints on the length of the Arabic composite name and our initial experimentation demonstrated high recall and precision results when the NER algorithm was applied to a financial domain corpus.

KEYWORDS

Arabic Named Entity Recognition; Natural Language Processing; Arabic language Grammar; Domain Knowledge

1. INTRODUCTION

Nowadays, an ever-increasing amount of information is available on the Web, covering a plethora of domains in business, education and entertainment. The overwhelming majority of this information is available in unstructured format. This limits the possibility of employing software technologies to process the information and extract new knowledge for the benefit of the end-user. Unstructured data refers to information that either does not have a predefined data model and/or does not fit well into relational tables; examples include email messages, word processing documents, Web pages and many other kinds of business document.

One of the approaches that are increasingly used to handle unstructured text is Natural Language Processing (NLP) [Rodrigues and Teixeira 2015], which is the core technology driving many applications such as Information Retrieval, Machine Translation, Data Mining and Question Answering [Ray and Shaalan 2016]. Named Entity Recognition (NER) is a fundamental task for NLP that is essential for extracting key terms related to a specific domain [Shaalan 2014]. Most reported Arabic NER efforts focus on techniques for extracting proper names in the text (e.g. persons, locations, and organisations) [Shaalan and Raza 2008][Alruily et al. 2014], but do not consider entities comprising composite names, i.e. entities identified by a concatenation of multiple words, where one or more of these words are names of a place, a person, an organization, etc.

Processing Arabic text is especially challenging compared to other languages such as English and European languages [Harmain et al. 2004] . This can be attributed to the fact that the Arabic language is morphologically rich, with short unwritten vowels and lacks capitalisation convention [Shaalan 2014]. Moreover, Arabic is orthographic with diacritics, and is highly inflectional and derivational. For example, the preposition or conjunctive may appear in one word as a prefix to the nominal, such as "لخدمات" (for services) or "وخدمات" (and services). In addition, the problem of extracting proper names is especially complex in the Arabic language, because the first letter of the word, which is capitalised in European languages, cannot be used to recognise proper names. Saad & Ashour in [Saad and Ashour 2010] and Shaalan & Raza in [Shaalan and Raza 2007] have mainly used indicator words, such as person indicators "الرئيس" (the president) or "الملك" (the King) or company indicator "شركة" (Company) to solve this problem. However, composite names can be composed of different phrases, such as place, or owner etc., and may contain several words, representing a mixture of nouns, adjectives and particles, which makes the automatic identification of Arabic composite names especially challenging.

In this work, we suggest a new approach for extracting Arabic composite names, which is based on the analysis of the definiteness and indefiniteness of composite names, in accordance with the Arabic grammar genitive principles. Part of Speech (PoS) tags are assigned for each word in the text, such as noun, verb, preposition, etc. We devised several syntactical rules to create linguistic patterns that match Arabic Genitive Rules (AGR) and composite organisation name patterns in financial and economic texts. We used recognition of Arabic organisation names as a use case, and the results of our initial experiments show high precision and recall scores.

The paper is structured as follows: Section 2 reviews related work; Section 3 discusses the architecture and implementation of our Arabic NER approach; Section 4 presents and evaluates the experimental results; and Section 5 summarises the paper and presents plans for further research work.

2. RELATED WORK

Various works have been published on Arabic NER. Shaalan [Shaalan 2005] presents the Arabic GramCheck system that addresses common grammatical errors occurring in Arabic language texts. The system is based on deep syntactic analysis, with reliance on a feature relaxation approach for detection of ill-formed Arabic sentences. It has two parts, namely an Arabic morphological analyser and a syntactic parser, extended to include a grammar checker. The system was evaluated using black box testing. The proposed approach detects grammatical errors using sentences as a unit of analysis, but does not propose how to address the detected errors. In addition, the corpus used to evaluate the approach is quite limited in size (100 sentences).

Zayed, El-Beltagy and Haggag [Zayed et al. 2013] present a new approach to Arabic person names recognition. A dictionary of Arabic named entity types was used to label Arabic names. They developed four rules to find the names of persons according to linguistic information about the names. The approach was applied in three domains, namely economics, sport and politics. The f-measure was used to compute performance in each domain, where the quoted scores were high, registering 92.04 for the economic domain, 92.66 for sport, and 90.43 for the political domain. However, the authors used Arabic dictionaries to recognise Arabic person names. This is a limited approach for some domains where it is not possible to provide dictionaries that contain an exhaustive list of names.

Oudah and F. Shaalan in [Oudah and Shaalan 2012] integrated rule-based and Machine Learning (ML) approaches to create a new hybrid approach to address Arabic NER tasks. Their approach was able to recognise 11 different types of Arabic entities, such as person, location, organisation, etc. Three different ML classifications were applied to evaluate the performance of the hybrid NER system. The authors report that the result outperformed the state-of-the-art of Arabic NER, in terms of accuracy, scoring f-measure results of 94.4 for person, 90.1 for location, and 88.2 for organisation. However, the paper does not detail how the integrated linguistic rules assist in captures composite names.

Elsebai, Meziane and Belkredim [Elsebai et al. 2009] developed and implemented systems to recognise person names in the Arabic language using a rule-based approach. The output of Buckwalter Arabic morphological analysis [Buckwalter 2004] was used as input to their system. They also used a set of keywords as indicators to phrases that contain a person name. The system was evaluated by comparing it with Person Name Entity Recognition for Arabic (PERA) [Shaalan and Raza 2007], which combines an initially collated NE lexicon (gazetteer) with a rule-based system that recognises the inflected form of the names. The authors state that the recorded precision, recall and f-measure registered 93%, 86% and 89% thus outperforming the PERA NER. However, their named enitiy recognition approach heavily relies on linguistic rules that are built specifically for Person name extraction, such as the introductory verb and word lists, a comprehensive list of names that start with letters "ال" (AL-the), and a dictionary list with *exclusion* words (city and country names); this makes the suggested approach difficult to readapt for recognising different types of entinities such as organisation names.

Traboulsi in [Traboulsi 2009] presents an approach that extracts Arabic NE by using local grammar rules, i.e. syntactic restrictions on certain sentences (e.g. verbal sentences) [Harris. 1991] to identify patterns of person names as function words clustered around Reporting Verbs (RV). Three analytical methods were used to find an Arabic person name: frequency, collocation, and concordance analyses. The author performed limited evaluation studies, which made it difficult to draw definitive conclusions about their achievement.

Wajdi Zaghouani [Zaghouani 2012] proposed an Arabic information extraction system called (RENAR) which aimed to extract different types of Arabic NE, such as person names, locations, organisations, date and numbers from different Arabic online news. The RENAR system relies on three main steps: pre-processing, lookup of full known names, and recognition of unknown names by using local grammars and a set of dictionaries. The authors reported that the system had performed well and provided good results with different Arabic named entities, except for the organisation category where the result was low due to several challenges, such as the extended length of the name and limited lexicon of the gazetteers.

Yassine Benajiba, Mona Diab, and Paolo Rosso in [Benajiba et al. 2009] experimented with the impact of using different sets of features (explore lexical, contextual and morphological) in three special machine learning frameworks, namely, Support Vector Machines (SVM), Maximum Entropy (MEnt)and Conditional Random Fields (CRFs) for the task of NER. Their work makes valuable contribution to the understanding of the impact of learning features on the Arabic NE task and the reported results show that the CRFs approach achieves a better result than Support Vector Machines and Maximum Entropy, but the paper does not explicitly analyse the fitness of the reported work for multi-token composite NER.

In [Omar and Al-Tashi 2018], Omar and Al-Tashi introduced a hybrid linguistic approach and a statistical method with a view to enhance the extraction of the Arabic nested noun compounds. The linguistic approach comprised part-of-speech tagging and named entities pattern, while the statistical method consisted of several association measures such as the combination-value, NLC-value, NTC-value, and NC-value. The authors reported that the performance of the combination-value is better than the other three association measures in terms of identifying Arabic nested noun compounds achieving accuracy of 90%. The proposed approach relies on a pre-constructed list of named entities to assist in identifying noun compounds, which is not suitable for domains where the composite NE tokens cannot be compiled in advance.

The authors in [Ali et al. 2018] present an NER approach for extracting several Arabic named entities such as people, location, organization and date. They use a neural networks approach and present a bidirectional long short-term memory

(LSTM) model that is claimed to achieve high accuracy for Arabic NER. However, similar to Benajiba et al. above, no particular focus on composite NEs was investigated in the paper.

A recent contribution in [Sayed et al. 2019] uses using a machine learning approach, which combines Conditional Random Fields (CRF) classifier and predefined gazetteers to improve NER accuracy. The approach is dependent on the availability of sufficient labelled data, hence the authors state that recognising NEs of organisations was poor “Due to the lack of tagged organizations in the training corpus”. It will be interesting to add the grammatical features (e.g. genitive patterns) to the feature matrix to investigate possible enhancement to their machine learning approach.

The reviewed works rely on lexicon-based (gazetteers), morphological analysis, and local grammar and machine learning to extract Arabic named entities, but without considering Arabic grammar in the syntactic analysis of text and most of the methods do not explicitly address the challenges of processing Arabic composite names. In this work, we investigate the rules of Arabic language grammar, in particular the rules regarding using the definite noun (معرفة) and indefinite noun (نكرة), to support the process of extracting composite names.

3. DEFINITENESS-BASED APPROACH FOR ARABIC NER

Our approach to extracting Arabic composite names from Arabic unstructured data is part of a larger effort that aims to investigate Semantic Web (SW) support for handling documents that are authored and/or annotated in Arabic [Khalil and Osman 2014]. The proposed approach comprises two methods, compilation of a PoS tag list that is used to remove the non-essential symbols and generate the PoS tag for each token in the Arabic sentence, and lexico-syntactic analysis for extracting Arabic composite names. The extracted entities are injected into a semantic knowledgebase, which forms the basis for document analysis of the next phase in our research. Fig. 1 shows the workflow of our Arabic NER approach.

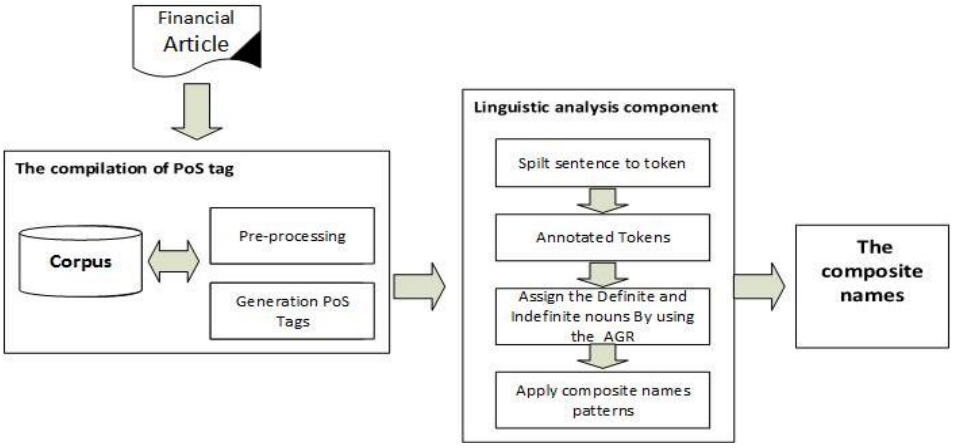

Fig. 1. Workflow of the Definiteness-based NER approach

3.1. Linguistic pre-processing

The first component of the system architecture is compilation of a PoS tag list. This involves removing non-essential symbols and generation of the PoS list for each token.

3.1.1. Removing Arabic variations and diacritics

Pre-processing is a way to improve the text by removing unnecessary elements. After converting our text corpus into UTF-8 encoding, in the first step, it is necessary to clean up the texts by removing punctuation marks, diacritics, numbers, non-Arabic letters. These pre-processing operations include:

- Removing punctuation that is attached to the word (+,- ,/,*,(,),_ ,&,%,.,),÷,×,],[,},{,<,>, \,|,@,#);
- Removing used diacritics and other signs (، ؛ ّ ْ ً ٌ ُ ٍ ِ َ);
- Normalising some writing forms that use different styles such as the optional inclusion of the letter “ء” (Hamza) as in “أحمد” (Ahmed) and “إمارات” (Emirates);
- Removing non-Arabic words for example “JVC”, “Vodafone”, etc.

Here we note that removing the diacritics can lead to the incorrect interpretation of the word’s semantics as in “أمينْ” (Ameen) which is a person’s name (proper noun) hence classified as Definite, while “امينٌ” (Amenoun) is translated as trustworthy (masculine adjective) and hence is Indefinite. However, considering that the overwhelming majority of documents published online are non-discretised, the impact is not significant for applications that chiefly process text from online material.

3.1.2. Tokenisation and Sentence Splitting

Initial tokenisation involves separating out individual words that are identified by the blank spaces or a special character between them; this is followed by splitting the document into individual sentences.

3.1.3. Part of Speech (PoS) tagging

Many processes are applied to the text to extract a PoS tag for each word using a PoS tagger, which assigns parts of speech to each word (token), such as noun, verb, and adjective [Rabiee 2011]. In this work, we used the Stanford PoS tagger that is based on the maximum-entropy model. It was originally developed for English at Stanford University [Green and Manning 2010], but now supports many languages, including Arabic, for which it claims 96.42% accuracy [Kanaan et al. 2005]. The Stanford PoS tagger was used to annotate each word with a PoS tag and saves the output in an initial list. This involved annotating more than 1300 documents to build a POS dictionary, which will be used during the processing of the ANER pipeline to assign the POS for each token in sentences. The POS dictionary consists of 21100 different words with each word assigned a POS. Due to the complexity of the Arabic language's morphology, some of the output of the PoS tagging process is erroneous; for the used corpus, the incorrect PoS tags represented 6%. Therefore, the initial output was reviewed by an Arabic language expert. The errors identified were encoded programmatically to improve the parser functionality.
Examples of errors in the PoS tagging are shown in Table I below.

Table I. Errors of Arabic PoS tagging

Examle1: للمصرف (For bank)	
Error	"ل" (preposition) "لمصرف" (NN)
Correction	"ل" (preposition) "لمصرف" (DTNN)
Example2: سهم (Share)	
Error	"س" (NN) "هم" pronoun
Correction	"سهم" (NN)

3.2. Basic Named Entity Recognition

3.2.1. Build domain vocabulary

Our knowledge-based approach assists the NER task in targeting specific entity classes that are of relevance to the problem domain; these classes result from the domain analysis that identifies the domain's key concepts and its relations. We use gazetteer lists in a dictionary-based approach to extract explicit named entities of the identified classes such as the names of companies "شركة" and shares "سهم" and currencies "عملة". The names of these classes are also used as indicator words to help recognise, via further linguistic processing, the named entities that are not explicitly mentioned in the gazetteer lists.

The scarcity of Arabic language resources was evident in the poorly populated gazetteer lists in the GATE (General Architecture for Text Engineering) NLP engine [Zaidi et al. 2010] that was utilised for our NER efforts; for instance, it had over 29K entries for City NE in English, 900 in Russian, but only 211 in Arabic. Hence, we resorted to public sources to enrich the gazetteer lists for the domain entity classes. Specifically, Maknaz (Arabic Thesaurus) [Maknaz.org 2001], which is an expanded Arabic resource with specialist list of descriptors or indexing terms integrated into an information system application, and also Linked Open Data (LOD), which refers to data published on the Web in machine-readable format [Bizer et al. 2011]. The most comprehensive LOD dataset is DBpedia, which is a community effort that aims to extract the structured information from Wikipedia for Information Extraction and Retrieval. Moreover, DBpedia contains more than 4.5 million entities and more than 3 billion triples for different languages and domains, such as country, city, etc. Although the Arabic version of the RDF is not available in DBpedia, the English version was used to extract the Arabic NE by using 'label property' (RDF: DBpedia: label), while the list of Arabic NEs has been reviewed manually. The LOD DBpedia dataset has been used to improve several kinds of gazetteers, such as country, city, organisation, person name and location. For instance, at the time of completing this study, an additional 10,000 city names were extracted from DBpedia to update the gazetteer list inherited from GATE.

3.2.2. Engineering of Arabic grammar rules for extracting Arabic Named Entity

The rule-based approach was used for basic NER; it is based on a set of human-crafted patterns to extract the named entities. In this study, a set of rules based on Arabic grammar was developed in order to extract Arabic NE. The rules were implemented using GATE's JAPE rule (JAVA Annotation Pattern Engine), which gives a finite-state transduction over annotations based on regular expressions. JAPE is a version of CPSL (Common Pattern Specification Language). The left-hand-side (LHS) of the rules consists of an annotation pattern description. The right-hand-side (RHS) consists of annotation manipulation statements. Annotations matched on the LHS of a rule may be referred to the RHS by means of labels that are attached to pattern elements [Zaidi et al. 2010] .

The main processing is carried out by gazetteer lists and a set of grammar rules. The JAPE rules are used to annotate the text and detect named (classified) entities, such as company name, stock market name, share name, etc. Fig. 2 illustrates JAPE rule for extracting the city named entity. In this rule, the token "مدينة" "(city) will be used as an indicator to annotate the next word. If the next word's kind of PoS is NNP or DTNNP, the system will recognise the phrase as a city name. The system will add several features to the city, such as: kind="city", rule="EX_CITY", category="NNP".

```
Phase:exe_date
Input:  word Token
//note that we are using Lookup and Token both inside our rules.
Options: control = appelt

Rule: EX_CITY
(
{ Token.string=="مدينة"}
({word.kind =="NNP"}|{word.kind =="DTNNP"}):City
 )
 :Tag
-->
:City.city={kind="city",rule="EX_CITY",category="NNP"},
:City.word={kind="city",rule="EX_CITY",category="NNP"}
```

Fig. 2. Screen shot of the JAPE rule to extract 'city' as named entity

The rule in Fig. 2 illustrates the case of annotating the Arabic NE (Named Entity) based on the list of gazetteers and a set of rules, where the PoS for each Arabic named entity is noun (NNP). In some cases, the Arabic NE appears in the text in different forms that do not match the words in the gazetteer list. For example, the name of the country in the example below appears in adjective form, which adds a suffix to the general noun. In Arabic, suffixes can indicate the gender and plurality of the noun as illustrated in Table II below.

Table II. Example showing the Country named entity in the text in adjective form

كشف وزير التجارة الجزائري الهاشمي جعبوب أمس الاول ان بلاده قررت استيراد 3 ملايين طن من الاسمنت في الاشهر القليلة المقبلة لتلبية حاجات السوق المتزايدة
Algerian Trade Minister El Hachemi Djaboub revealed yesterday that his country has decided to import three million tons of cement in the following few months to meet the growing needs of the market.

In the sentence in Table II above, the country name ("Algerian"," الجزائري") is used as an adjective word by adding the suffix letter ("ي") to the original word ("الجزائر"," Algeria"). Therefore, the Jape rule illustrated in Fig. 2 was extended to remove the suffix and match the resulting token against the Country gazetteer list.

Determining the token boundary of an Arabic named entity is considered one of the main challenges for Arabic NER [Shaalan and Raza 2009][Alanazi 2017]. Some Arabic NE's are complex names that are composed of different phrases and may also contain several words, representing a mixture of nouns, adjectives and particle, which makes the automatic identification of Arabic composite names more challenging. There are some studies that have attempted to solve this problem by using gazetteer lists [Zaghouani 2012], but dictionary-based approaches are not suitable for recognising entities in continuously updated lexicon such as organisation names. Based on Arabic genitive grammar rules, this study presents a novel approach that uses domain knowledge, i.e. the problem domain's key concepts and relations, to formalise a set of syntactical rules and linguistic patterns to extract Arabic composite names from unstructured texts.

3.3. Linguistic analysis for composite name extraction

Our approach involves two main stages, the first one is to classify the words as definite or indefinite nouns, and the second is pattern recognition to extract composite names. We devised two pattern recognition mechanisms, the first uses Genitive Patterns for classifying definiteness within phrases, and second develops Linguistic patterns to extract composite names.

3.3.1. Grammar-based analysis to classify words as definite or indefinite

Grammar-based analysis is applied to classify pronouns into definite noun (الإسم المعرفة) and indefinite noun (الإسم النكرة), which is a key stage in our approach to extracting composite names. The definite noun is one that refers to a specific noun (person, animal, thing, etc.) such as "محمد" (Mohammad) / "الشركة" (the company). The main Arabic grammar rules for definiteness and indefiniteness as explained in Table III below.

Table III. List of the grammar rules for definiteness and indefiniteness

Types of Definite Nouns	Example	English translation
The Proper Noun (refers to a specific name of someone or something or someplace) <اسم العلم>	محمد	Mohammed
Definite Noun identified by the definite article (the) < الإسم المعرف بـ (ال) >	البيت	the house
Possessive Pronoun (their) < الضمير (هم) >	سيارتهم	their car
Relative Nouns (that) < الأسماء الموصولة (الذي) >	العالم الذي اخترع	scientist that invented

Demonstrative Nouns < أسماء الإشارة (هذا) >	هذا المصنع	this factory
Indefinite Noun added before a definite noun < المضاف الى معرفة >	لخدمات الحاسوب	for computer services
The Object of Vocation	يا رجلان	Hey two men

In the last Definite type in the table above: "لخدمات الحاسوب" (for the computer services), the word "لخدمات" (services) is indefinite, but it is considered definite as it is added to the proper noun "الحاسوب" (the computer) which defines which "خدمات" (services) is meant.

In Arabic grammar, the indefinite noun (الإسم النكرة), is one that refers to a common and non-particular noun (person, animal, thing, etc.). It can be given to any member under that category of nouns. For example, "مدينة" (a town); "شارع" (a street); "دولة" (a country). Based on the above, we devised rules to classify the tokens annotated at the previous stage into definite and indefinite nouns, based on the following conditions. Table IV shows the abbreviations that are used in the this work based on the Stanford tagger.

- If the kind of token is DTNN or DTNNS or DTJJ or DTJJS or NNP, then the token will be identified as DE.
- If the kind of token is NN, then the token will be identified as INDE.
- If the kind of token indicates non-noun; for example, verb or preposition, the system will reject this token. We also needed to devise new rules to classify the tokens associated with genitive articles, such as preposition and conjunction, as explained below.
  - If a preposition is used, such as in "لخدمات" (for services), where the word "خدمات "(services) is combined with the preposition "ل" (for), then the word "خدمات" will be identified as indefinite. However, since it is combined with a preposition, it is identified as (INDEIN).
  - Where a conjunctive (حرف عطف) "و" (and) is used to join two or more tokens together, such as in " لخدمات وصيانة" (for services and maintenance), our system will classify this phrase as in the above explanation. So "لخدمات" (for services) will be classified as INDEIN. Moreover, since the word "وصيانة" (maintenance) is combined within the previous word using a conjunction, it will also be identified as Indefinite (INDECC).

Table IV. Reference list of PoS tags

Symbol	Description	Symbol	Description
DT	Articles including 'a', 'an',	DTJJ	adjective with the determiner "Al" (ال)
IN	preposition	DTJJS	a plural adjective with a definite article attached
JJ	adjective	NN	noun - singular or mass
DTNN	noun, singular or mass with the determiner "Al" (ال)	NNP	proper noun
DTNNS	noun, plural with the determiner "Al" (ال)	NNPS	proper noun – plural
NNS	noun – plural	INDE	Indefinite
INDEIN	Indefinite word attached with preposition	INDECC	Indefinite word attached with conjunctive
DE	Definite	CC	conjunction: "و" (and)

3.3.2. Pattern recognition to extract composite names

This section explains the development of linguistic patterns that are used to retrieve composite names from the unstructured text where the nouns were classified as definite and indefinite as detailed in the previous section. Our approach uses two types of patterns to extract the information. The first is used to construct phrases based on Arabic Genitive Rules. The second pattern is used to extract the composite name.

3.3.2.1. Genitive Patterns for classifying definiteness within phrases

The first pattern (the 4th (last) row in Table V below) describes the syntactic sequence used to extract the *definite* phrase that contains an indefinite word "خدمات" **preceded** by conjunctive "و" (CC - حرف عطف) and succeeded by a definite word "النظافة"; Arabic Genitive Rules (AGR) rules are then applied to tag the phrase as definite DE.

Table V. Example illustrating the mechanism of the first Genitive Pattern (GP1)

Example	Construction and cleaning services company - شركة الانشاءات وخدمات النظافة				
Tokenisation	النظافة	خدمات	و	الانشاءات	شركة
	the cleaning	services	and	The construction	Company
PoS tagging	DTNN	NN	CC	DTNN	NP
Classification pattern	DE	INDECC		DE	Indictor
Definite tagging	DE			DE	Indictor

The second pattern deals with sentences that contain propositional phrases. As illustrated in Table VI, AGR is applied to the classification pattern in the table to define the propositional phrase "لخدمات" (for service) as definite by associating it with the succeeding Definite noun "الحاسوب".

Table VI. Example illustrating the mechanism of the second Genitive Pattern (GP2)

Example	Garyounis Company for Computer Services شركة قاريونس لخدمات الحاسوب				
Tokenisation	الحاسوب	خدمات	ل	قاريونس	شركة
	the computer	service	for	Garyounis	Company
PoS tagging	DTNN	NN	IN	NNP	NP
Classification pattern	DE	INDE	Preposition	DE	Indictor
	DE	Prepositional Phrase		DE	Indictor
Definite tagging	DE			DE	Indictor

The third pattern above can be extended to include word(s) that join the propositional phrase with the conjunctive "و" (حرف عطف). The classification pattern in Table VII illustrates how the two indefinite words "صيانة" (INDECC) and "برمجة" that joined the propositional phrase "لخدمات " have also been added to the definite classification.

Table VII. Example showing the mechanism of the third Genitive Pattern (GP3)

Example	شركة قاريونس لخدمات وصيانة وبرمجة الحاسوب Garyounis Company for Computer Services, Maintenance and Programming								
Tokenisation	الحاسوب	برمجة	و	صيانة	و	خدمات	ل	قاريونس	شركة
	the computer	programming	and	maintenance	and	service	IN	Garyounis	Company
PoS tagging	DTNN	NP	CC	NN	CC	NN	IN	NNP	NN
Classification pattern	DE	INDE	Conjunction	INDE	Conjunction	INDE	Preposition	DE	Indictor
	DE	INDECC		INDECC		Prepositional Phrase		DE	Indictor
Definite tagging	DE							DE	Indictor

Now that each token has been correctly identified as a definite noun " الاسم المعرفة" or indefinite noun "الاسم النكرة", taking into consideration the genitive rules at phrase level, the next stage is to apply patterns that were devised to extract the actual composite names.

3.3.2.2. Linguistic patterns to extract composite names

One of the greatest challenges in Arabic NER is the lack of capitalisation for proper nouns. This challenge is commonly addressed by using indicator words relevant to the problem domain in order to narrow down the search space for ca corresponding named entity. In this work, we also use indicators, which are referred to as trigger words within our patterns and are used for marking name phrases that might contain a corresponding domain-relevant name such as "company" (شركة) that aids to identify the name phrase "Arabian Gulf Oil Company" (شركة الخليج العربي للنفط) word locating named entities and their semantic meanings in unstructured text relevant to our targeted economic domain.

For extraction of Arabic composite names from the unstructured text, we use two linguistic patterns that take into consideration the attachment of the definite article "ال" (the) to the Arabic composite names. The first pattern considers indicators that do not have a definite article attached, and the second pattern considers indicators that have a definite article attached.

The first pattern is used to extract the composite names when indicated with "ال" (the), such as "المؤشر" (the share) indicator quoted in the example below. In this pattern, to construct the composite name, the consecutive definite (DE) words/phrases, succeeding the indicator word will be added to the composite name until an indefinite (INDE) word, such as "تداول" (Trading) as in Fig. 3 below, is encountered.

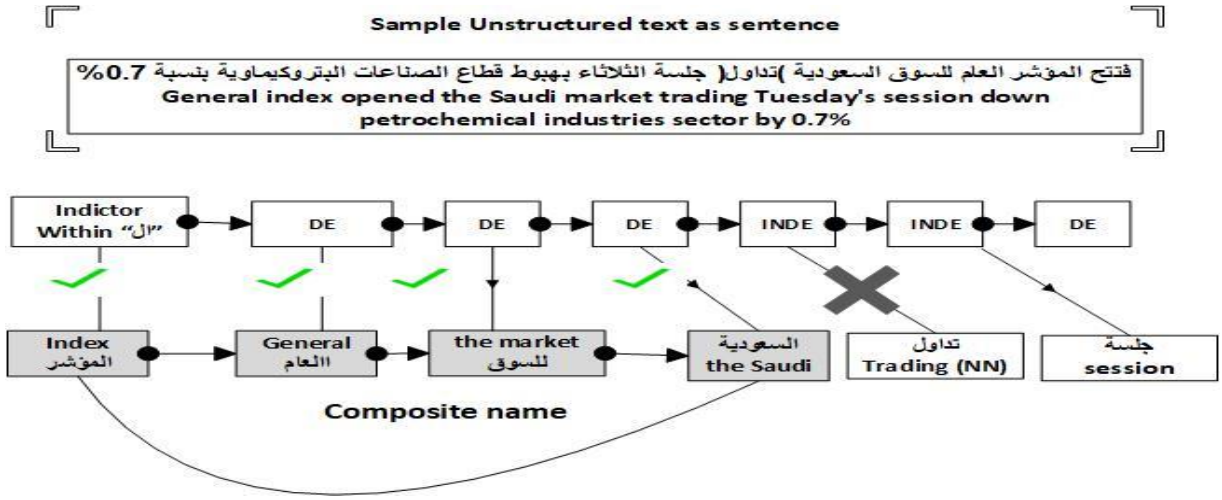

Fig. 3. The mechanism of composite name extraction using the first Linguistic Pattern (LP1)

The second pattern is used to extract the composite names that are not attached to the definite article "ال" (the), such as "قطاع" (Sector) in the text exemplified in Fig. 4 below. In this pattern, the word immediately following the indicator may be either DE or INDE. However, similar to the previous pattern rules, all the consequent words must be of type DE, as illustrated.

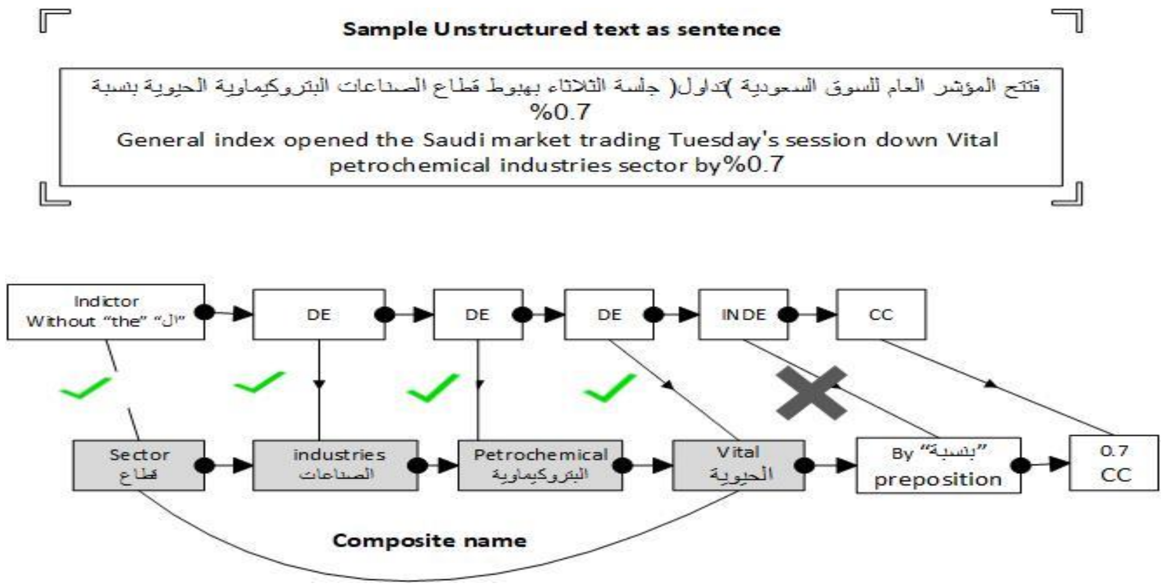

Fig. 4. The mechanism of composite name extraction using the second Linguistic Pattern (LP2)

The algorithm below explains the steps to extract the organisation's composite names by using genitive processing of an organisation name extraction pattern.

Algorithm1: Implementation of the linguistic analysis for composite name extraction (in pseudocode)

```
START
  Input: Raw Text, Indicator Words, Arabic Genitive Patterns (GP1, GP2, GP3), Linguistic Patterns (LP1, LP2)
  //Stage I – identify definite nouns
  Perform initial pre-processing: Tokenisation, Sentence Splitting, and Part of Speech Tagging
  for each sentence in the pre-processed text
    for each token in the sentence
```

```
      if (the PoS of the token = (NNP OR DTNN OR DTNNS OR DTJJ ) then
        token kind = DEFINITE noun
      elseif (The PoS of token = NN or NNS or JJ) then
        if (the first letter in token = "ل") then
          The kind of token = INDEIN
        elseif (token = "و" ) then
          if the length. Token >3
            The kind of token = INDECC
          endif
        else
          The kind of token = INDEFINITE noun
        endif
      endif
    endfor
  endfor
  //Stage II – identify definite phrases
  for each sentence in the text
    use the indicator word to find domain-relevant phrases
    use the Genitive Patters GP1, GP2, GP3 to identify the DEFINITE subphrase
  endfor
  //Stage III – Extract composite names
  for each domain-relevant phrase
    if the indicator word is DEFINITE //LP1
      continue to add subsequent DEFINITE word or DEFNINITE phrase to the composite name
      until end of phrase OR INDEFINITE word is encountered
    else //indicator word is INDEFINITE – LP2
      add next word to composite name whether DEFINITE OR INDEFINITE
      continue to add subsequent DEFINITE word or DEFNINITE phrase to the composite name
      until end of phrase OR INDEFINITE word is encountered
    endif
  endfor
  Output: Composite Named Entities
END
```

4. SYSTEM EVALUATION

This section documents the evaluation of the composite name extraction algorithm. The aim is to evaluate the performance of our NER system in extracting composite Arabic names with varied complexity in terms of length and grammatical structure.

4.1. Data Collection

At the first stage of evaluation, we collected a set of digital newspapers related to our case study (from the economic domain). The text corpus was collected from different websites that represent different authoring styles. For instance, documents published on websites related to the stock market domain have a special (bulletin – type) style compared to other general economic news websites. More than 1000 news articles were collected to evaluate our approach. Table VIII and Table IX show the contribution of the different online sources to the test data and the corpus specification of the collected documents. Manual annotation was performed for the document corpus, across all targeted entities. Even though this process is time consuming, it results in improved precision. The precision recall and f-measure were analysed in two experiments, the first experiment evaluates the effect the length of composite names and the second experiment compares the results of using different genitive patterns.

Table VIII. Documents' collections

Resources	Share of test data
http://www.fxnewstoday.ae/	13%
http://sa.investing.com/	12%
https://www.icn.com/ar/	06%
http://www.aljazeera.net/ebusiness	07%
http://www.alborsanews.com	25%
http://www.bbc.com/arabic/business	08%
Other news	31%

Table IX. Corpus specification

Item	Quantity
Resourses	9
Document	1300
Sentences	6055
Tokens	189290
Composite names (CN)	4710
- Two words CN	- 1242
- Three words CN	- 1960
- Four words CN	- 1066
- More than four CN	- 442

4.2. Discussion of the results and limitations

At the time of compiling this paper, we could not find any published research with public datasets and results that evaluated NLP efforts at extracting Arabic named entities comprising composite names for a specific problem domain. Hence, an evaluation could not be directly compared against published efforts in the field. The experiment aimed to evaluate the performance of our system in extracting the composite names dependent on number of words in the composite names and the AGR within the composite name. During the first experiment, we observed that the names, which contain two words demonstrated better performance in terms of precision (95.6%), as shown in Fig. 5, compared to three words (93%), four words (91.5%) and five or more words (92.5%). Hence, while we can note a slight decline in precision as the number of composite words increase, the developed entity recognition system maintains high accuracy.

In a few cases, the definiteness-based pattern recognition can lead to false detection of composite named entities. For example in the phrase: "مليون سهم تقريباً بالجلسة الماضية" (the last session has almost a million shares), in which the word following the indicator "سهم" (share) is the indefinite word "تقريباً" (almost). So, the definiteness-based pattern recognition system incorrectly determined that the "سهم تقريباً" (almost share) is a composite name.

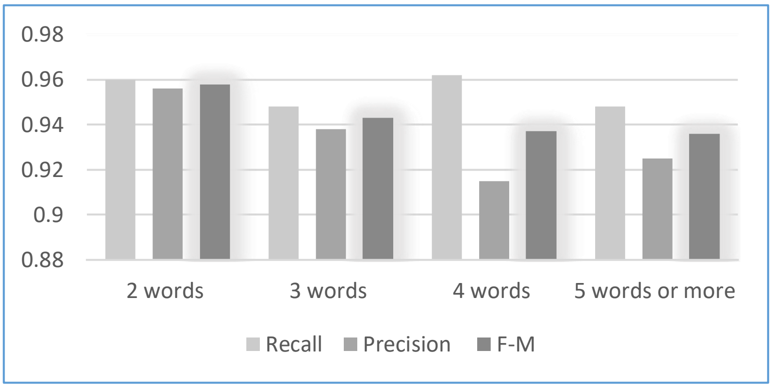

Fig. 5. Impact of composite names' length on precision

The second experiment compares the use of different AGR patterns on composite name extraction. Arabic composite names can take different forms as illustrated in in Table X. The experiment evaluates the genitive NER patterns devised in section 3.3.2. as mentioned above. As illustrated in Fig. 6, we observe that the names that used the first pattern recorded higher precision (100.0%) compared to the more structurally complex second (95% precision) and third patterns (96% precision) with consistent high recall and F-measure results. Therefore, we can claim that the genitive pattern method proved its consistency and accuracy for the recognition of Arabic composite names.

Table X. Examples of the Arabic composite names recognised with AGR patterns

Patterns	Arabic composite names example	English trans
First	شركة التصنيع وخدمات الطاقة	Industrialisation and Energy Services Company
Second	شركة العربي الافريقي الدولي لتداول الاوراق المالية	Arab African company for financial trading
Third	سهم قاريونس لخدمات وصيانة وبرمجة الحاسوب	Garyounes company for service, maintenance and computer programming

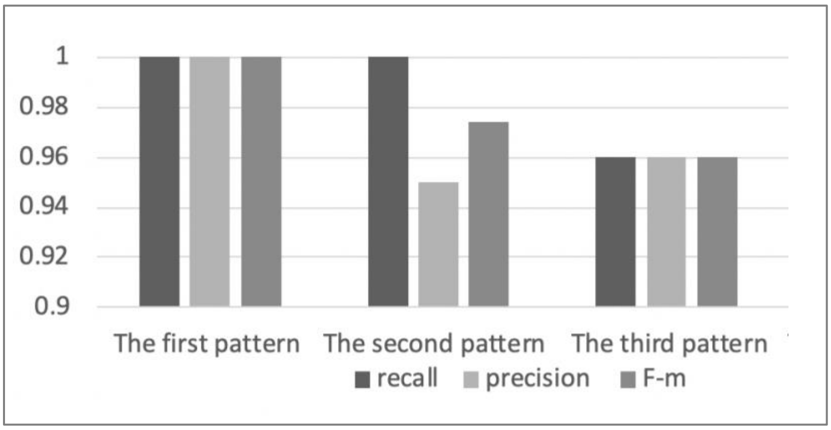

Fig. 6. Recall, precision, and F-measure of the AGR patterns

It is noteworthy that the analysed results were affected by problems associated with the shortcomings of the PoS tagger, as well as grammatical mistakes in the original text. The deployed Stanford tagger erroneously tagged some words as nouns, although they were verbs. It also incorrectly tagged the word “سهم” as “س+ هم”; it is separated into the proclitic (س) and (هم) as a pronoun. Similarly, the PoS Tagger currently cannot detect the proposition lam “ل”, which can affect the accuracy of the composite NER in case the word following the indicator word starts with the lam “ل” letter; for instance, the word “لحية” (beard) can be parsed as “ل+حية” (with+snake). It also had problems with words translated from foreign language such as “جي في سي” (JVC) Although we have a mechanism to manually handle such exceptions as explained in section 3.1.3 above, however, where there is a significant impact on the information retrieval accuracy, a more systematic approach is required to fundamentally manage the above explained errors.

Some syntactic analysis errors were caused by grammatical mistakes in the authored text, such as “شركة اسمنت ابيض” (White Cement Company). The rules of the Arabic language do not allow three words or more to be joined together to compose an indefinite Type (نكرة). Our approach cannot deal with these names, because the base of our algorithm does not allow three or more words to be joined together in an indefinite manner.

5. CONCLUSIONS

Arabic NER, especially for composite names, is a challenging task due to the complex morphology of the Arabic language and lack of advanced Arabic NLP tools. In this paper, we presented a novel approach for extracting composite names from documents authored in the Arabic language. Our approach to composite Arabic entity recognition is based on the genitive grammar rules of the Arabic language where we initially use grammar-based analysis to classify pronouns into definite noun (الإسم المعرفة) and indefinite noun (الإسم النكرة). We then devised a set of genitive pattern recognition rules to retrieve composite names from unstructured text; the first set of rules uses Genitive Patterns to classify definiteness within phrases. and the second set of linguistic patterns uses Arabic definite articles to extract the composite names from the classified phrases.

Experimental evaluation was performed on financial documents with varied authoring styles and revealed good precision and recall results. It also confirmed that our error correction mechanism applied to the output of the PoS tagging process results in noticeable improvement in the effectiveness of our composite names extraction approach. The paper also highlights unresolved problems relating to the complex Arabic PoS tagging process, and to syntactic analysis errors stemming from common misuse of the Arabic language grammar. The next stage of our research will focus on relation extraction to capture events that are of relevance to the chosen domain of interest. This will complete, together with recognised named entities, the semantically tagged data that will be injected into a knowledgebase providing for the intelligent exploration of Arabic unstructured documents.

As information retrieval challenges include unstructured text that is difficult to associate with specific domain, our future plans for future work involve investigating the utilisation of advanced Machine Learning techniques for NER [Yadav and Bethard 2018] in *generic* Arabic text that is not associated with a particular problem domain.

REFRENCES

RODRIGUES, M. AND TEIXEIRA, A. 2015. Advanced applications of natural language processing for performing information extraction. Springer, .

RAY, S.K. AND SHAALAN, K. 2016. A review and future perspectives of Arabic question answering systems. *IEEE Transactions on Knowledge and Data Engineering* 28, 3169-3190.

SHAALAN, K. 2014. A survey of Arabic named entity recognition and classification. *Computational Linguistics* 40, 469-510.

SHAALAN, K. AND RAZA, H. 2008. Arabic named entity recognition from diverse text types. In *Advances in Natural Language Processing,* Anonymous Springer, , 440-451.

ALRUILY, M., AYESH, A. AND ZEDAN, H. 2014. Crime profiling for the Arabic language using computational linguistic techniques. *Information Processing & Management* 50, 315-341.

HARMAIN, H.M., EL KHATIB, H. AND LAKAS, A. 2004. Arabic text mining. In *IADIS International Conference Applied Computing,* Anonymous , 23-27.

SAAD, M.K. AND ASHOUR, W. 2010. Osac: Open source Arabic corpora. In *6th ArchEng Int. Symposiums, EEECS,* Anonymous .

SHAALAN, K. AND RAZA, H. 2007. Person name entity recognition for Arabic. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources,* Anonymous Association for Computational Linguistics, , 17-24.

SHAALAN, K.F. 2005. Arabic GramCheck: A grammar checker for Arabic. *Software: Practice and Experience* 35, 643-665.

ZAYED, O.H., EL-BELTAGY, S.R. AND HAGGAG, O. 2013. A Novel Approach for Detecting Arabic Persons' Names using Limited Resources. *Research in Computing Science* 70, 81-93.

OUDAH, M. AND SHAALAN, K. 2012. A pipeline Arabic named entity recognition using a hybrid approach. *Proceedings of COLING 2012* 2159-2176.

ELSEBAI, A., MEZIANE, F. AND BELKREDIM, F.Z. 2009. A rule based persons names Arabic extraction system. *Communications of the IBIMA* 11, 53-59.

BUCKWALTER, T. 2004. Buckwalter Arabic Morphological Analyzer Version 2.0, Linguistic Data Consortium (LDC) catalog No LDC2004L02. 2019, .

TRABOULSI, H. 2009. Arabic named entity extraction: A local grammar-based approach. In *Computer Science and Information Technology, 2009. IMCSIT'09. International Multiconference on,* Anonymous IEEE, , 139-143.

HARRIS, Z. 1991. Theory of language and information: a mathematical approach.

ZAGHOUANI, W. 2012. RENAR: A rule-based Arabic named entity recognition system. *ACM Transactions on Asian Language Information Processing (TALIP)* 11, 2.

BENAJIBA, Y., DIAB, M. AND ROSSO, P. 2009. Arabic named entity recognition: A feature-driven study. *IEEE Transactions on Audio, Speech, and Language Processing* 17, 926-934.

OMAR, N. AND AL-TASHI, Q. 2018. Arabic nested noun compound extraction based on linguistic features and statistical measures. *GEMA Online® Journal of Language Studies* 18, .

ALI, M., TAN, G. AND HUSSAIN, A. 2018. Bidirectional Recurrent Neural Network Approach for Arabic Named Entity Recognition. *Future Internet* 10, 123.

KHALIL, H. AND OSMAN, T. 2014. Challenges in Information Retrieval from Unstructured Arabic Data. In *UKSim,* Anonymous , 456-461.

RABIEE, H.S. 2011. Adapting standard open-source resources to tagging a morphologically rich language: a case study with Arabic. In *Proceedings of the Second Student Research Workshop associated with RANLP 2011,* Anonymous , 127-132.

GREEN, S. AND MANNING, C.D. 2010. Better Arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics,* Anonymous Association for Computational Linguistics, , 394-402.

KANAAN, G., AL-SHALABI, R. AND SAWALHA, M. 2005. Improving Arabic information retrieval systems using part of speech tagging. *Information Technology Journal* 4, 32-37.

ZAIDI, S., LASKRI, M.T. AND ABDELALI, A. 2010. Arabic collocations extraction using Gate. In *Machine and Web Intelligence (ICMWI), 2010 International Conference on,* Anonymous IEEE, , 473-475.

MAKNAZ.ORG. 2001. Maknaz - Expanded Arabic Thesaurus. 2018, .

BIZER, C., HEATH, T. AND BERNERS-LEE, T. 2011. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts,* Anonymous IGI Global, , 205-227.

SHAALAN, K. AND RAZA, H. 2009. NERA: Named entity recognition for Arabic. *Journal of the American Society for Information Science and Technology* 60, 1652-1663.

ALANAZI, S. 2017. A named entity recognition system applied to Arabic text in the medical domain.

GALICIA-HARO, S.N., GELBUKH, A. AND BOLSHAKOV, I.A. 2004. Recognition of named entities in Spanish texts. In *Mexican International Conference on Artificial Intelligence,* Anonymous Springer, , 420-429.

YADAV, V. AND BETHARD, S. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics,* Anonymous , 2145-2158.

SAYED, A.M., ABDOU, S., RASHWAN, M. AND AL-BARHAMTOSHY, H., 2019. RANER: RDI Framework for Arabic Named Entity Recognition. *International Journal of Engineering & Technology,* 8(1.11), 161-164.