International Journal of Applied and Natural Sciences (IJANS)
ISSN (P): 2319-4014; ISSN (E): 2319-4022
Vol. 7, Issue 2, Feb – Mar 2018; 65-86

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

A PROPOSED MODEL FOR EXTRACTING INFORMATION FROM ARABIC-BASED CONTROLLED TEXT DOMAINS, DISCUSSING THE INITIAL MODEL STEPS

Mohammad Fasha, Nadim Obeid & Bassam Hammo

Research Scholar, Department of Computer Science, King Abdulla II School for Information Technology, The University of Jordan, Amman

ABSTRACT

Information extraction from Arabic as well as other languages text is commonly implemented over restricted text domains. Approaching open text domains is challenging, because of the syntactic, semantic and pragmatics ambiguities and variations in text. For the purpose of approaching more relaxed versions of Arabic text domains, Fasha et al. (Fasha et al. 2017) presented a high-level description fora proposed work methodology that can establish a model for extracting information from controlled text domains. In that work, controlled text domains were defined as the text domains that are not restricted in their linguistic features or their knowledge types yet they are not very unanticipated in these respects. In this paper, we discuss that work methodology and its implementation in more detail. Our discussion includes the initial phases of the methodology which covers the corpus preparation processes including its selection, analysis and annotation using a custom morpho-syntactic Part-of-Speech tagging scheme, we also discuss the designing of the supporting knowledge-base model which will be used to represent and process the extracted information. The information extraction algorithm itself shall be presented in a future work.

***KEYWORDS**: Narabic Natural Language Processing POS Tagging Ontology Based Information Extraction Description Logic*

Article History

***Received:** 13 Jan 2018 | **Revised:** 30 Jan 2018 | **Accepted:** 03 Feb 2018*

1. INTRODUCTION

Information Extraction (IE) from text is usually implemented on restricted text domains where the targeted information and the category of text are tightly constrained. This restriction is forced by the need to confine natural language variations to minimize the syntactic, semantic and pragmatics ambiguities it mighten compass. For that reason, approaching more relaxed versions of text domains is challenging and efforts are rarely presented in these areas.

Information extraction from Arabic text has additional challenges, these challenges are caused by some of Arabic's distinguishing features which includes its rich morphology, highly inflectional nature, free word order, subjects dropping and the omitting of diacritics or short vowels in Mean Standard Arabic (MSA). These distinguishing features increase the challenges for disambiguating text while the correct interpretation is inferred from the context (Attia 2006; Sawalha et al. 2013).

In an attempt to mitigate some of the previously mentioned challenges, Fasha et al. (Fasha et al. 2017) presented a high level description for a proposed work methodology that can establish a model for extracting information from more-relaxed versions of Arabic text domains. In that work, these relaxed versions of text domains were identified as controlled text domains which were defined as the text domains that are not restricted in their linguistic features or their knowledge types yet they are not totally unanticipated in these respects. The objective of that multi-step work methodology was to examine all the key processes that might be involved in transforming Arabic text into formal knowledge representations. Another objective of that work methodology was to present a framework that can be employed on different types of controlled text domains while addressing all the involved keys steps for establishing the IE model and not making any key assumptions for neutralizing any of the common challenges.

In this work, we examine that work methodology in more detail. Our discussion covers the initial phases of the proposed methodology which includes the selection of a suitable corpus, the analysis and the annotation of the corpus and the designing of the supporting knowledge-based model.

The contributions of this work include presenting the new concept of controlled text domains and we approach children stories as a type of controlled text domains, and to the best of our knowledge this is one of the few works that addresses IE from Arabic-based children stories, which is an important area that might assist in setting basis for approaching other types of relaxed text domains. We also discuss a custom Arabic Part-of-Speech (POS) tagging scheme that can better address Arabic's distinguishing features including its rich morphology and its well-established declension system. Further, we present an event-based ontology that can represent the main concepts and relations that are found in short narratives such the selected text domain. Finally, we introduce a Java-based implementation that can be used to annotate Arabic text with morpho-syntactic POS tags as well as the designing of supporting ontologies.

The remaining of this paper is structured as follows. In section 0we discuss some of the related previous work. Section 0 presents the work methodology which we followed in this work. Section0presents a brief introduction about the new concept of controlled text domains. In section 0, we describe the corpus preparation process including its selection, analysis and annotation. In section 0, we discuss the designing of the knowledge representation model and finally the conclusion and suggested future work are presented in section 0.

2. RELATED WORK

In this work, we discuss a model related to extracting information from relaxed versions of Arabic text domains and we selected children stories as a type of text domains to be placed under investigation. Reviewing previous work revealed a scarcity in work related to the extraction of information from stories in general and children stories in specific. Nevertheless, work is occasionally introduced in related areas such as the work that was presented by (Kim et al. 2016) who proposed a model for establishing basic understanding about news stories from a timeline-summarization perspective. In that work, the authors argued that establishing an understanding about news stories based on the dynamically changing relations between entities is simpler than reading the whole set of text blocks. To assess that assumption, the authors presented a rule-based model that can establish Subject-Verb-Object (SVO) triples for the temporal relations between the main events in text. The model was evaluated on FIFA 2014 World Cup news articles and the results were found promising according to the authors.

Another similar work was introduced by (Samson and Ong 2014) which presented a model for extracting conceptual binary relations from children stories. The objective of that work was to investigate a model that can automatically establish semantic relations that can be employed later for automatically generating children stories. The proposed model utilized a rule-based scheme based on using GATE platform and custom extraction rules to extract a set of predefined binary relations. The model was evaluated on (30) normalized children stories and it yielded low accuracy (36%) on the precision, recall and f-measure scales. The main cause for the low accuracy of the extracted information was justified by relating it to the over generalized and simplified extraction patterns.

In a similar area, (Wang and Zhao 2012) discussed an ontology-based model that can better represent the different types of information that are related to news events while focusing on representing knowledge that can answer the(5W1H) types of questions, i.e. what, who, when, where, why and how types of questions. To evaluate the proposed ontology, a prototype Chinese news story was manually represented to assess the capability of the designed ontology to represent all the required information.

Reviewing related Arabic literature on story understanding, we were able to identify the single of (El-Salam et al. 2013). Which discussed a graphical model that can describe the animations in a story. That work was mainly comprised of general discussions about the challenges that might be involved in extracting information from Arabic text with some highlights on proposed schemes that can be applied to address these challenges. The authors proposed using event calculus to represent and process the extracted knowledge, but no evaluation or results were presented to demonstrate the implementation of that proposed model.

In brief, no thorough model for extracting information from children's stories were found, most of the examined work focused on a specific area of that domain, more importantly, its representation using an adequate knowledge representation model.

As for Arabic-based information extraction work, most of the presented models implemented rule-based schemes where the extraction rules were manually defined. This fact might be justified by the scarcity of proven and coherent Arabic-based supporting tools and resources, including the different NLP libraries as well as the labelled corpora which are required to enable the different machine learning based approaches. A relatively recent work in that area include the work of (Omri et al. 2017)which presented a model for answering questions from a temporal perspective. That model was based on searching for specific Arabic time-related keywords e.g. (when, day, date, since…etc.) that signal the presence of temporal information related to events and consequently extract information that can answer time-related questions.

Similarly, (Mesmia et al. 2017) presented a rule-based model that employed finite state transducers to extract semantic relations between Arabic Named Entities. (Bentrcia et al. 2017) presented a rule-based model for enriching the construction of Quranic ontology based on Arabic conjunctive patterns.

In the same respect, (Sadek and Meziane 2016a) discussed a model that can detect explicit causal relations from Arabic text. That model was based on employing a rule-based pattern-recognizer that can detect the cause-effect relations in text. The recognizer operated based on around (700) linguistic patterns that were manually extracted based on a manual analysis process on an untagged Arabic corpus. In (Sadek and Meziane 2016b) a model for answering questions based on the work of (Sadek and Meziane 2016a) was presented. The objective of the model presented in (Sadek and Meziane 2016b) was to answer nonfactual types of questions i.e. why, how vs. the factual based questions what, who and where that are related to a given Arabic narrative.

A model for extracting information related to future events from newspaper articles was presented in(Alruily and Alghamdi 2015). The main theme of that work was to employ Arabic future verbal proclitic particles i.e. (س sa, will or سوفsawfa, will) to signal and identify future events in a "pre-normalized" set of news articles.

A model for extracting terms from special Arabic corpora was presented by (Al-Thubaity et al. 2014). That model employed an algorithm that was based on two assumptions; the first one was that terms are frequently used in special domain corpora and the second assumption was that special domain terms are bounded by special word types e.g. prepositions, determiners and conjunctions or by orthographic signs such as numbers or punctuations.

A model for the automatic extraction of ontology relationships from Arabic text e.g. the cause-effect, is-a, part-whole, has-a, kind-of relations was presented by (Al Zamil and Al-Radaideh 2014). In that work, (Hearst 1992) algorithm was employed to implement a lexical-semantic based extraction process.

In contrast to the previously discussed models, the model that we are implementing work does not make any assumptions for neutralizing any of the tasks that are commonly involved in extracting information from text, rather, all the key steps that might be involved in the text-to-formal knowledge transformation are addressed. In addition, the domain of interest in our study and the types of the knowledge types that we are targeting are more varied that the previously discussed models, this includes explicit and implicit concepts and relations as well as atomic and composite ones. Finally, the model we are implementing propose to employ Ontology Web Language (OWL) ontologies in practical implementations using Java applications while exploitation the inferencing capabilities of Description Logics (DL) to further extend the extracted knowledge.

In the next section, we present a high level description about the proposed work methodology that was initially introduced in (Fasha et al. 2017)and we discuss its initial steps in more detail.

3. THE PROPOSED WORK METHODOLOGY

The work methodology that we are implementing in this work was initially introduced in (Fasha et al. 2017). The motivation for proposing that detailed word methodology - presented in Figure below- was driven by the scarcity of previous similar work for Arabic language. Hence, it was important to define a systematic work methodology that can investigate all the key steps that might be involved in implementing and evaluating a model for extracting certain types of information from a newly approached text domain. A main objective of this detailed work methodology was to avoid making any assumptions for neutralizing known challenges that usually involve extracting information from Arabic text.

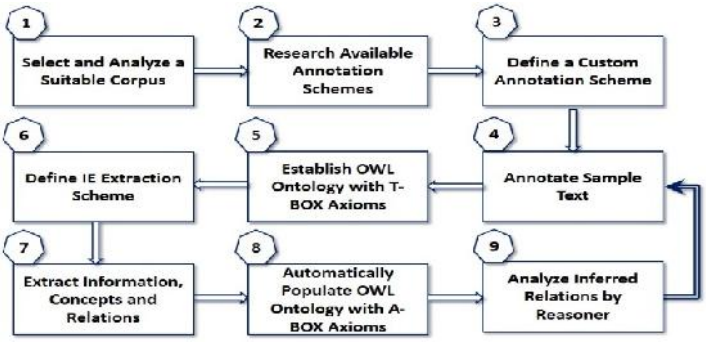

Figure 1: The Proposed Work Methodology

The implemented work methodology resembles a pipeline of progressive steps where each step receives input from a previous one and delivers processed output to the next phase of processing.

The proposed work methodology is mainly comprised of two main parts, the first one is related to the preparation of the corpus and the knowledge base supporting model and the second one is related to the implementation of the proposed information extraction scheme.In this respect, (Fasha et al. 2017) proposed implementing a two phase information extraction process where the first phase targets atomic types of information that are explicitly mentioned in text while the second phase shall extend that initially extracted knowledge by employing the inference capabilities of Description Logics.

In this work, we concentrate on the initial preparation phases of the proposed methodology (steps 1 to 5) which include the identification of a compatible text domain, the selection of a corpus in that domain, analyzing that corpus looking for main syntactic and semantic features and annotating the corpus using a suitable POS tagging scheme. In addition, the initial preparation steps involve the designing of the supporting knowledge base supporting model that can represent all the initially extracted knowledge and extend that knowledge by inferring new types of implicit and composite concepts and relations. The outcomes of these initial phases provides the basis that are needed to establish the proposed IE scheme later.

In the next sections, we provide a brief discussion about the concept of controlled text domains which are targeted by the proposed work methodology and later we discuss the work we implemented to fulfill the model preparation phases.

4. INTRODUCING CONTROLLED TEXT DOMAINS

Commonly, information extraction work in Arabic language is implemented over specific types of text domains where the vocabulary and terminology, syntax and semantics are well defined and constrained e.g. medical, social, legal, technical text domains etc. In addition, these conventional implementations usually target a specific and reduced set of information that can serve a specific objective e.g. the recognition of named entities, identifying future events, extracting causal relations, looking for specific keywords…etc.

In an attempt to approach more relaxed types of text domains from a wider perspective in terms of information extraction, (Fasha et al. 2017) introduced the concept of controlled text domains which were defined as the text domains that are not constrained in their syntax, semantics and pragmatics yet they are not totally unanticipated in these respects.The objective of introducing this new concept is to present a generalized definition for a new category of text domains that falls between the restricted text domains and the open free texts.

The items listed below presents some of the distinguishing features that can be identified about controlled text domains:

- In controlled text domains, the diversity of the targeted information is more relaxed and wider than the ones that are usually targeted in restricted text domains.
- The type of targeted knowledge includes the information that is explicitly mentioned in text as well as the implicit ones that represent background knowledge. Also, the targeted information might include atomic concepts and relations as well as composite and higher order ones.

- The objective of the information extraction from a controlled text domain is definable. This definition can assist and serve in establishing general understanding about a coherent narrative vs. single sentence understanding.
- The vocabulary, syntax, semantics and pragmatics in controlled text domains are not constrained. Each new narrative can introduce new occurrences for any of these features.
- Although the vocabulary, syntax, semantics and pragmatics might change, yet in controlled text domains there exists some clear guidelines or patterns that can assist in establishing adequate representations for that knowledge and consequently establishing rules that can extract these types of information.

Several domains can fall within this definition of text domains; this includes the different types of narratives that discuss life facts or events such as the domain we selected in this study i.e. younger ages children stories.

5. CORPUS PREPARATION

In this section, we discuss the corpus preparation process in more detail. The fact that there was no previous similar work in Arabic language and following the proposed work methodology, we commenced our work by preparing a corpus that can be used to implement and evaluate the proposed work methodology and the proposed IE model.

The corpus preparation process includes the identification of a compatible text domain, the selection of a suitable corpus within that domain, the analysis of the corpus to identify main syntactic and semantic features and the annotation of the corpus using a suitable Part-of-Speech tagging scheme. More information about these tasks is presented below.

5.1: Domain Identification, Corpus Selection and Analysis

Several types of textual domains can be classified under the presented definition of controlled text domains and one of them is children stories, which we selected as a candidate text domain to be placed under investigation in this work.

Children stories were selected, because they encompass different types of knowledge and their syntactic features and semantic content are not restricted, each story can discuss a different moral or theme whereas new words and concepts can arise without restrictions. In addition, the selection of children's stories was based on the assumption that their syntax, semantics and pragmatics content are relatively simpler than elder ages open text since they target younger age's audience. Besides, stories in general and children stories in specific were identified as an important stepping-stone for approaching wider and more relaxed types of text domains(Riloff and Riloff 1999). Furthermore, introducing work on children stories was compelling because up to our knowledge, this is one of the few Arabic-based works that approaches children stories in IE implementations. Finally, according to the early work of (Madar and Sarmistha 1989), understanding a story can be defined as the ability to answer objective and subjective questions about the story and to be able to summarize that story at different levels. This succinct and concise definition was used to guide our different efforts that targets the establishing of the proposed IE model.

Reviewing Arabic literature for a compatible corpus, we were able to identify (Seyoufi 2014)younger ages children story series as an adequate corpus that can serve our objectives. That story series was comprised of (24) short narratives that are grouped into five categories where each category convoys a certain moral or theme e.g. manners and behavior at school, with friends, with family, with parents and with neighbors.

Two main tasks were performed on the selected corpus. First, the corpus was manually analyzed to identify the general syntactic and semantic features that are incorporated in text, and second, that corpus was annotated using a custom (POS) tagging scheme as shall be presented later. The objective of analysis process was to identify the key features that can assist in providing generalization guidelines which can be employed during the designing of the knowledge representation model and the IE scheme later.

Table 1next presents a summary about the main findings of the corpus analysis process. From a linguistic perspective, it was found that most sentences in the corpus were verbal sentences or sentences that were centered around verbs. In addition, most sentences encompassed subjects whether they were explicitly mentioned or dropped. Another main finding was that even for a relatively minimal corpus like the one we selected, it was noticed that the linguistic structures were rich and variant and involved various syntactic and grammatical features of Arabic language. For example, for any given story, we observed occurrences for different verb tenses, subject dropping, co-referencing as well as free word order and the role of rich Arabic morphology and its word derivations and inflections were present.

Table 1: Main Findings about the Selected Corpus

Item	Count
Total Number of Stories	24
Total Number of Sentences	280
Average Story Sentences	11
Verbal Sentences	221
Nominal Sentences	59
Nominal Sentences with Verbs	50
Nominal Sentences without Verbs	9
Sentences with Verbs	271
Sentences with references to Characters	272
Sentences without references to Characters	8
Verbs with pronouns	Present
Nouns with pronouns	Present
Subject dropped	Present
Assertive Sentences	88%
Imperative Sentences	10%
Interrogative and Exclamatory	2%

On the other hand, from a semantic perspective, it was observed that most sentences were discussing events or activities that were related to story characters, and each story encompassed a set of atomic concepts and relations that build up to compose more-general concepts that can convey a certain moral or theme to its audience.

Figure 2 next presents a pseudo representation of the implicit and composite knowledge types that were observed during the analysis process. The illustrated assertions describe the actions that might be performed by a certain character as well as the consequences of these actions.

Samples for Atomic and Explicit Relations
Character1 did Action1 Action1 caused Action2 Action2 caused Action 3 Action3 is Bad Actions are transitive → Action1 caused Action 3 Character of Action1 is the real Character of Action3
Samples for Composite and Implicit Relations Bad Action → Character1 is a bad person Doing bad Action and Regretting → Character1 is Good Person Good is inverse to Bad → Character1 is not a Bad Person Character did or caused some bad action and felt regret → Good Character

Figure 2: General Description of the Targeted Composite and Implicit Knowledge

For example, a character might perform an event that causes another event, which results in some damaging effects, later, that character confesses his actions, and acknowledges their consequences, apologizes and demonstrates regret. Therefore, a conclusion can be reached that such a chain of events is tutoring children not to do such regrettable events and to apologize when they do so, which also might imply that the character is a good mannered individual. Different stories can be composed around this general principle or theme. In the same manner, a character might be involved in a chain of events that ends in a positive outcome, and thereby educate children about good deeds.

The general finding concerning the semantic part of the analysis process complies with the notions that were presented in (Samson and Ong 2014). In that work, the knowledge types of children stories were classified into main categories, the operational knowledge that defines the sequencing of story plots and the background knowledge which provides information about story characters, the world and the causal relations between events.

The general findings about the corpus analysis process we reemployed later to define the main features of the corpus annotation scheme as well as the during the designing of the knowledge-supporting model. More information about these tasks are provided in the next sections.

5.2 Corpus Annotation

The model that was presented in (Fasha et al. 2017)proposed to implement the information extraction process in two phases where the first phase targets atomic concepts and relations while the second one targets implicit and composite concepts and relations. The implicit and composite knowledge is inferred based on the findings of the first phase of information excretion as well as the background knowledge incorporated in the supporting knowledge base. Hence, this high dependency between the first phase of the IE process and the second phase of IE indicates that failing to accurately extract a given atomic concept or relation might affect the correct extraction of the targeted implicit and composite concepts and relations later.

As proposed in (Fasha et al. 2017), the first phase of the IE process shall be implemented using regular expressions and Part-of-Speech (POS) tagging which will be emplyed to extract atomic and explicit information from text, therefore, it was important to use a rich POS tagging scheme that can enable the accurate representation and extraction of information.

In this respect, the results of the analyses process that was presented in the previous section revealed several challenges related to Arabic's linguistic features in the selected corpus. This includes occurrences for subject dropping cases, free words order and the use of detached and attached pronouns. For example, the sentence (وركل كرته فاتجهت نحو الزجاج فحطمته, wrkl krth fitajahat naHowAlzjAj fahatamathu, and he hit his ball so it went towards the window and it broke it) has three dropped subjects, the character that kicked the ball and the ball itself that went towards the window and broke it. In addition, an enclitic is attached to the noun "ball" which references the subject i.e. possession pronoun. Another attached pronoun is present in the verb (فحطمته, so it broke it) which indicates a singular masculine object referring to the glass. Consequently, it was concluded that an adequate morphology-aware Part-of-Speech tagging scheme was needed to correctly and accurately POS tag our selected corpora.

In the next section, we discuss the work that we implemented to identify a suitable POS tagging library using the available and accessible related resources as well as the resolution that we followed in regard.

5.2.1: Examining Existing Resources

In an attempt to identify a suitable morphology-aware Part-of-Speech tagging library, we examined a number of the available and accessible POS taggers and morphology analyzers for Arabic language. The tools we examined included Stanford NLP toolkit (Manning et al. 2014), NLTK toolkit (Bird 2006), AL-Khalil morphology analyzer (Boudlal et al. 2010), BAMA morphology analyzer (Buckwalter 2002), MADAMIRA (Pasha et al. 2014) and SAFAR platforms (Souteh and Bouzoubaa 2015).

To run the examination, we selected a number of Arabic sentences and we analyzed using these different resources. Table below presents the main findings of the performed examinations for one of the examined sentences.

Table 2: The Results of the POS Tagging Examination

Arabic Sentence	وركل	كرته	فاتجهت	نحو الزجاج	فحطمته
English Translation	and he hit/inflected to Singular Masculine Subject Inflection	his ball/ attached possession pronoun	So it went Singular Feminine Subject Inflection	towards the window/glass	so it broke it Singular Feminine Subject + Singular Masculine Object
Buckwalter Transliteration	wrkl	krth	fitajahat	naHow AlzujAj	fHTmt
Stanford NLP	وركل VBD /	كرته/NNP	فاتجهت/VBD	نحو/NN الزجاج/DTNN	فحطمته /VBD
Al-Khalil	12 solutions, verbs and gerunds	17 solutions, verbs, nouns and gerunds	5 solutions, verbs	15 solutions for both	13 solutions, verbs and gerunds
BAMA	2 solutions including VERB_PERFECT and NOUN	6 solutions, NOUN	6 solutions including VERB_PERFECT With different subject inflections	4 solutions for naHowa and 5 solutions for AlzujAj	9 solutions VERB_PERFECT Different Subject and Object inflections

Table 2 Contd.,					
Arabic Sentence	وركل	كرته	فاتجهت	نحو الزجاج	فحطمته
SAFAR MADAMIRA	7 solutions, different subjects inflections	15 solutions, different subjects and objects inflections	4 solutions, different subject inflections	14 solutions for naHowa and 17 solutions for Al zujAj	12 solutions, different subjects and objects inflections
MADAMIRA	wa/CONJ+rakal/PV+a/PVSUFF_SUBJ:3MS	kur/NOUN+ap/NSUFF_FEM_SG+a/CASE_DEF_ACC+hu/POSS_PRON_3MS	fa/CONJ+{it~ajah/PV+at/PVSUFF_SUBJ:1S	naHowa/PREP Al/DET+zujAj/NOUN+a/CASE_DEF_ACC	fa/CONJ+HaT~am/PV+at/PVSUFF_SUBJ:3FS+hu/PVSUFF_DO:3MS
Custom Tagging	({RP+WA+CC} {VBD+SNG+MSC+3rd})	({NN+SNG+FEM}{SUFX_POSS+SNG+MSC})	({RP+FA+CZ} {VBD+SNG+FEM+3rd})	({RB+LOC})({NN+MSC})	({RP+FA+CZ} {VBD+SNG+FEM+3rd}{SFX_OBJ+SNG+MSC)

As demonstrated in the table, a number of challenges that can affect the automated annotation process were observed. For example, Stanford NLP (POS) tagger produced a set of over generalized POS tags since it uses the reduced English Treebank tag set. In addition, the resultant POS tags of that tagger were not morphology aware. Moreover, numerous errors were witnessed in the resultant POS tags such as incorrectly the following words as proper nouns e.g. (وكلن, and he was; بها, with it;وقال له , and he said to him…etc.).

Similarly, BAMA and AL-Khalil well-known rule-based morphology analyzers generated all the possible interpretations for a given word while the user is responsible for identifying and selecting the correct interpretation according to the context. Moreover, Al-Khalil library generates its results in plan Arabic text according to Arabic declension system, which will require significant customization and processing to be able to employ its outcomes in establishing POS based extraction patterns.

SAFAR framework was also experimented and found capable and can be integrated with other modules since it is based on Java language. On the other hand, the framework was a collection of other tools i.e. BAMA, AL-Khalil, MADAMIRA that were combined into the unified framework of SAFAR. Accordingly, this resource did not present any unique or additional features other than the ones provided by the original libraries. Similarly, NLTK provided support for training and running POS taggers but it is also uses English Treebank tag set and it was not morphology aware.

MADAMIRA toolkit was also investigated and found compelling and capable, and similar to SAFAR it is based on Java, which can be easily integrated into different implementations. Moreover, this tool is context-aware which implies that words are analyzed and tagged according to their role in the sentence. Nevertheless, MADAMIRA also produced some erroneous tags on different occasions, for example, the verb (فاتجهت, so it went) was inflected as a first person singular masculine rather than a third person singular feminine while some other words were not resolved at all. Such errors or short comes can affect the accuracy of the sought POS tagging scheme.

Another important observation that was noticed about the examined resources was that they could generate overlapping between the characters of the generated POS tags. Such duplications or overlapping can cause complications when employing any string matching mechanisms such as regular expressions.

For example, the sample tags presented below demonstrate how the same single character overlaps between an atomic unit and a composite one e.g. ***F*** for Feminine and ***F*** as part of the ***SUFF*** keyword.

*VERB_**P**A**SS**IVE+**PVSUFF**_**S**UBJ:3FS*

*VERB_**P**A**SS**IVE+**PVSUFF**_**S**UBJ:3**FS***

*VERB_**P**A**SS**IVE+**PVS**UFF_**S**UBJ:3M**P***

Similarly, it was observed that the same concept could be represented using different markers within the same scheme. For example, in the sample tags presented below, we can observe that singular number was represented using the ***SG*** marker in the first sample and using an ***S*** character in the second. The same is true for the feminine gender marker.

*ADJ+NSUFF_**FEM_SG***

*IV3**FS**+VERB_IMPERFECT*

In this respect, it was determined that it is more convenient if we can use a consistent marker for eachsinge concept that we need to represent in a POS tag in order to facilitate the definition of the matching patterns later.

Finally, the examination process of the related resources revealed that there was no standardized scheme for the POS tagging process; some libraries employed general POS tags while others employed morphology aware tags. Similarly, there was no standardized POS tokenization scheme and different formats might be used by the different libraries to aggregate the composite tag markers. Consequently, different tag set were used by the different libraries, some were generalized while others were more specific.

In conclusion, all the examined libraries included some short comes and manual intervention was required to correct these errors or to fine-tune the results to incorporate any missing information. Consequently, it we defined a custom POS tagging scheme that can facilitate the later processes of the proposed model. The purpose of that scheme was to address the distinguishing features of Arabic language and it should be sufficiently accurate and precise. In the next section, we present the custom POS tagging scheme that was prepared as well as the custom Java module that was prepared to assist in implementing that scheme and annotating the selected corpus.

5.2.2 The Custom POS Tagging Scheme

Based on the findings of the examination process that was presented in the previous section, we prepared a custom annotation scheme that can support a “fine-tuned” Arabic compliant POS tagging process. The main objective of that scheme was to allow users to commence with an Arabic compliant POS annotation process in a clear, simple and agile manner.

In principle, the designed POS tagging scheme was based on consolidating the tokenization of words along with the POS tagging process. This consolidation was achieved by employing different types of brackets which were used to establish word level groupings of the word morphemes and the designated POS tags.

Two types of brackets were using within the custom scheme, the round brackets or parenthesis “( )” and the braces or the curly brackets “{ }”. The parentheses were used to define word boundaries while the braces were used to create aggregates or sub-tokens of the morphemes within each word.

The listing presented below demonstrates the POS tagging of the surface word (wa sa nokhberu hum, وسنخبرهم, and we shall inform them) according to the proposed custom scheme:

- The proclitic morpheme (wa, و, and, coordinating conjunction) is annotated as {RP+WA+CC}.
- The proclitic morpheme (sa,س, shall, indicates a future event) is annotated as RP+SA+FTR}.
- The inflection particle (nun, ن indicates first voice plural speaker) is annotated as {PLRL+stV}.
- The stem (khabara,خبر , tell, the verb itself) is annotated as {VB}.
- The enclitic morphemes (hum,هم , them, an attached pronoun that indicates a plural masculine object) is annotated as {PRN+SFX_OBJ+PLRL+MSC}.

The advantages of the proposed scheme over the previously investigated ones are the following:

- The format of the proposed tagging scheme is customizable; it can support the generalized formats such as the one used by Stanford library as well as the detailed ones which are employed by the morphology aware libraries such as MADAMIRA. The user has the ability to customize the tagging details according to his needs while distinguishing words segments using brackets.
- The user can commence with the annotation process in an agile and seamless manner. The tokenization prerequisite is eliminated as the user can perform tokenization and POS tagging in the same time using the custom scheme and its enabling tool.
- No single characters were used to convey meaning; rather, each individual concept e.g. gender, number, tense…etc. is distinguished using a specific tag that can be aggregated with other tags using the plus sign '+' character. This feature preserves the clarity of the meaning and it can facilitate any string-based matching operations.
- Having such dynamicity and simplicity in defining markers and aggregating them using the brackets allow users to seamlessly introduce any specific makers that might be required to serve a specific objective, a plug-in-and-play manner of operation. This includes syntactic based markers, morphology related ones or any other types of markers such as special markers that might be related to Arabic language and its unique declension system.

To enable the annotation process, we prepared a custom Java-based module[1]which can be used to annotate any Arabic language corpus using the proposed POS tagging scheme. In this respect, we also clarify that our initial survey revealed that there is no available or accessible similar tool for Arabic language.

This tool – presented in Figure 3 below – can be used to accelerate and guide the corpus annotation process. The tool starts it operating by accepting a delimited text file where it will automatically normalize words and POS tag them using Stanford POS tagger.

[1]All the source code for the presented modules is available at https://sourceforge.net/projects/arabicie/

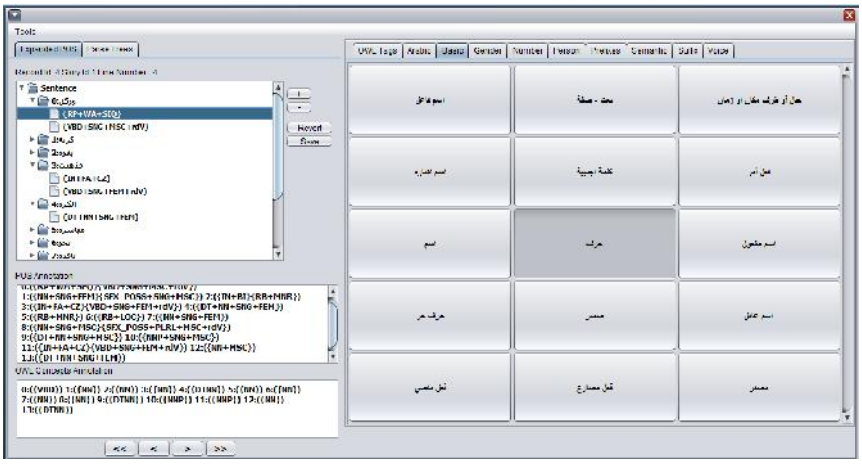

Figure 3: Screenshot of the Custom POS Annotation Tool

Next, these automatically generated POS tags are presented to the user using an intuitive Graphical User Interface (GUI) where he can validate, correct or extend any tag in a simple and seamless manner. Furthermore, each word is automatically distinguished using a unique number that can maintain the tag-to-word mapping, which can be employed within the information extraction algorithm later. In addition, a regular expressions friendly format is generated for each sentence where Regex control characters e.g. (“( ) { } : +”) are replaced with neutral characters that are suitable for Regex-based matching operations.

This tool was used to annotate sample stories from the selected corpus and it demonstrated competent results. Table 3 next presents annotation for a sample sentence using the custom annotation scheme. The outcome of the POS tagging pocess is presented in the original format (column 2) as well as the Regex neutral format (column 3).

Table 3: Annotating a Sample Story Sentence Using the Custom POS Annotation Scheme

Sentence	Custom POS Tagging	Regex Friendly Version
0:اهدت 1:ليلى 2:شقيقها 3:تامر 4:كره 5:جديده، Layla gave her brother tamer a new ball as a gift	0:({VBD+SNG+FEM+rdV}) 1:({NNP+SNG+FEM} 2:({NN+SNG+MSC}{POSS+SNG+FEM} 3:({NNP+SNG+ MSC}) 4:({NN+SNG+ FEM}) 5:({JJ+SNG+FEM})	<«VBD_SNG_FEM_rdV»> <«NNP_SNG_FEM» <«NN_SNG_MSC»«POSS_SNG_FEM» <«NNP_SNG_MSC»> <«NN_SNG_ FEM»> <«JJ_SNG_FEM»>

As presented in the sample sentence, the proposed annotation scheme is a combination of conventional syntactic POS markers as well as morphology related ones. The generated string for each sentence includes the words along with their unique identification number while the composite words are segmented according to their constituent morphemes and every segment is grouped using the curly brackets and the complete word is surrounded by a pair of parenthesis. In addition, the user can use the custom annotation tool to incorporate new markers or define new formats that might not have been defined by previous tools e.g. add gender markers for proper names, which can be useful in experimenting different information extraction scenarios.

In the next section, we present the work that was implemented to define the knowledge supporting model that can be used to represent the different types of concepts and relations about the selected children stories as well as inferring new types of knowledge.

6. KNOWLEDGE REPRESENTATION MODEL

A main feature of the proposed information extraction model that was initially introduced in (Fasha et al. 2017) was to employ Description Logic to assist in enhancing the information extraction process.

Description Logics are a family of formal knowledge representation languages that are used to represent and reason about concepts and relations in a certain domain and they provide logical formalism for different ontologies languages including Ontology Web Language (OWL)(W3C OWL Working Group 2012).The main building blocks of DL are concepts, roles and individuals. The relations between concepts, roles and individuals are defined using formal logical statements called axioms, which are divided into two categories: the terminological axioms or (T-Box) axioms and the Assertion Axioms or (A-Box) axioms. Terminological axioms are used to define the fixed facts about a certain domain while the assertion axioms are used to define instance values or individuals that are instantiated under the different types of concepts that are defined in the ontology.

The designing of knowledge representation models involves the definition of the taxonomy of the concepts and relations i.e. the (T-Box) axioms, that can describe facts and rules about a certain domain of interest. The designing of ontologies can vary according to the domain of interest as well as the objectives of the ontology (Zhou et al. 2004), therefore, we employed the design guidelines that were presented by(Uschold and King 1995) to establish the required ontology, these guidelines involve the following tasks:

- Identify the purpose and the scope of the ontology.
- Identify the key concepts and relations for the domain of interest.
- Identify the actual terms.
- Coding the ontology.
- Placing that design under evaluation.

The purpose of the knowledge-supporting model was to enable the information extraction process from the selected controlled text domain i.e. children stories. More precisely, the sought ontology should be able to incorporate knowledge that can answer the (5W1H) questions i.e. What, Who, Why, Where, When, and How, about the main events of the story as well as identifying the main themes or morals of that story. Hence, the scope of the ontology was bounded to the extent that can serve that purpose.

To enforce guidelines ii to v, we followed a custom approach that was based on employing Protégé along with a custom tool that we prepared to facilitate the designing process of the ontology. Mainly, the applied method involved examining the selected corpus sentence by sentence to identify concepts and relations and to define an OWL compliant representation for them. The objective of that detailed process was to identify and represent as much information as possible in order to create a highly accurate information extraction process later.

The purpose of the custom tool demonstrated in Figure 4 next was to eliminate the need to iterate back and forth between the text and protégé since doing so turned to be cumbersome especially when the ontology grew in size. The custom tool presents a view for both the text and the OWL representations in the same screen where the can perform any modifications directly using this tool.

To illustrate the working of scheme, we consider the sample sentence (اهدت ليلى شقيقها تامر كرة جديدة. *Layla gifted (gave a gift) her brother Tamer a new ball*). This sentence includes the following explicit concepts and relations:

- Event: Give.
- Agent of the event: Layla.
- Event First Object: Tamer.
- Event Second Object: The Ball.
- Property of the ball: New.
- The brother-sister relation between Layla and Tamer.
- The time of the giving event, which is a past tense in this case.

Accordingly, the custom tool was used to draft prototypes of these concepts and relations, including their taxonomy and naming, and to assure that no conflicts or overlapping are created with any previously defined concept or relation.

Later, we used the same tool to label words in sentences with their semantic roles. This labeling process established a golden-rule dataset that can be employed later for validating the accuracy of the information extraction process.

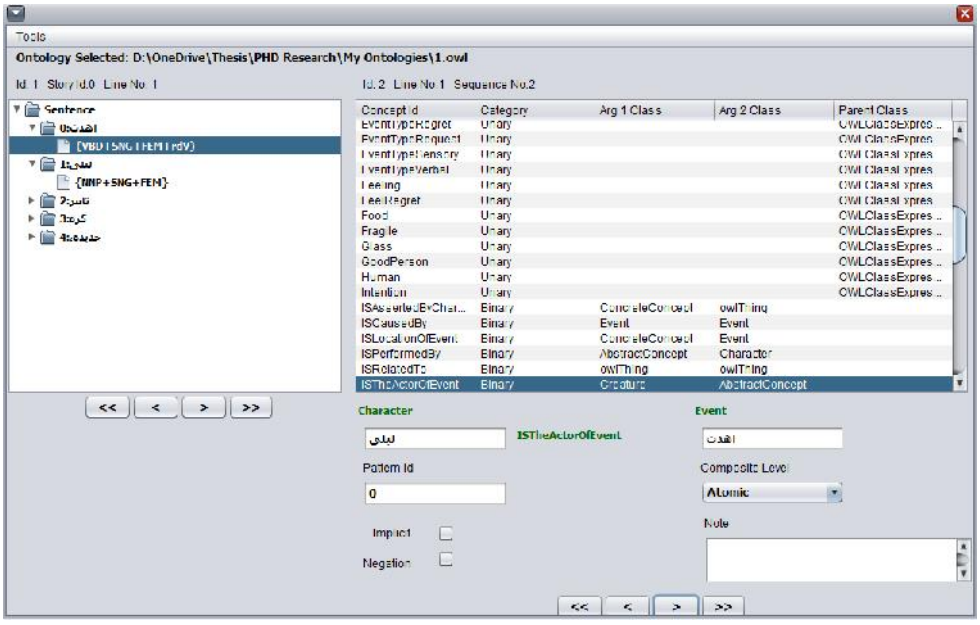

Figure 4: The Custom Tool for Mapping Text to OWL

A main consideration that we maintained during the designing of the ontology was to keep the defined concepts and relations as general as possible and not to create any specific concept or relation unless it affects the flow of the story. This direction was important in order maximize the potentials of sharing the generated ontology between the different stories.

For example, if the previous concept (give) does not affect the flow or the theme of the story, It can be instantiated by the information extraction algorithm under the generalized concept class e.g. *Event Concept*. On the other hand, if that event does influence the flow of the story or if it exhibits important implicit knowledge, then it was defined as a unique class under the general *Event Concept* class and it is augmented with additional specialized assertions that can better serve the inference process. As an example for this situation, if a given story is woven around the concept of lying, and the event of lying had some consequences within the context of the story, then that event is defined as a specific concept *Event Lie* under the concept *Event Concept.* In addition, that concept can be defined as a subclass of another concept that identifies its uniqueness by indicating the type of knowledge, it conveys e.g. *Bad*in order to explicitly provide the required additional knowledge.

Using the method briefed above, we concluded to the general taxonomy demonstrated in Figure 5 below. As demonstrated in the diagram, the core concepts of our selected corpus were related to events and characters as well as the set of concepts and relations that are tailored around these two main concepts. This taxonomy complies with the initial findings that were presented during the corpus preparation process that was discussed in section0.

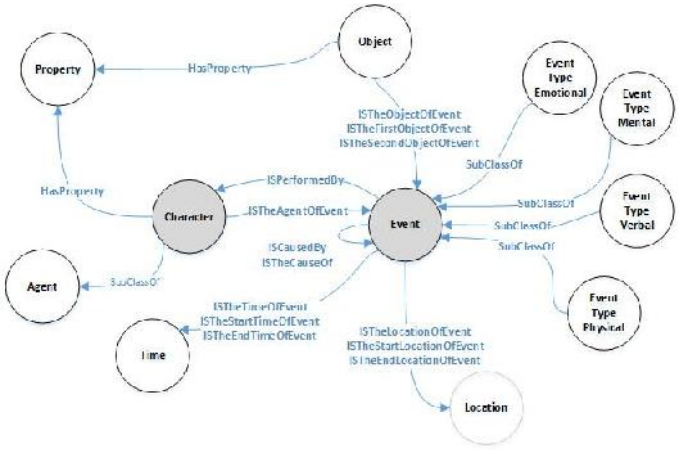

Figure 5: Event Based Ontology

Mainly, the designed ontology was comprised of the following three main categories of concepts:

a. Atomic Concepts

Atomic concepts or unary classes were used to represent the concrete and abstract concepts that were explicitly mentioned in text i.e. world knowledge. Concrete concepts include the physical objects that were referenced in stories e.g. *Boy, Ball, Window*…etc. while abstract concepts include information about different types of events that were mentioned in stories. In this respect, events were classified into the following categories according to their semantic relatedness:

- Physical events: Events that have tangible physical consequences in real world e.g. {(ضرب, hit), (اعطى, give) }.
- Sensory events: Events that are related to the character's sensory apparatus e.g. {(شاهد, see), (سمع hear) }.
- Verbal events: Utterance made by characters. This includes character's demands, assertions, requests, and inquiries e.g. {(قال said), (سأل Asked) }.
- Emotional events: Events that exhibit character emotions e.g. {(شعر felt), (خاف, Feared) }.

- Mental State events: Events that express the hypothetical state that corresponds to the thoughts, feelings or intentions of story characters e.g. {(اعتقد, thought), (قرر decided) }.

All actions mentioned in the stories were defined under the general concept of *Event* or one of its sub classes. This is true for deliberate and a deliberate event, including characters' actions, feelings, thoughts and emotions which were all considered as events according to the designed ontology. This taxonomy was found simple and capable of representing all the types of event-based information that was incorporated into stories.

b. Binary Relations

Binary relations include the roles that define relations between different concepts, which includes the relations between events and their different constituents, more importantly the type of the relations that answer the (who, what, why, when and where) types of questions.

The following list presents the most important binary relations that were defined according to the observed facts in selected story corpus:

- *IS The Agent Of Event (Agent, Event)*: Represent the agent or the story character that performed a certain action or event.
- *IS the Object Of Event, IS The First Object Of Event, IS The Second Object Of Event (Object, Event)*: Represent the object/s of events i.e. transitive verbs.
- *IS The Location Of Event, IS The Begin Location Of Event, IS The End Location Of Event (Location, Event)*: Represent the location information of events.
- *IS The Time Of Event, IS The Begin Time Of Event, IS The End Time Of Event (Time, Event)*: Represent the temporal information about events.
- *IS The Manner Of Event (Manner, Event)*: Represent the manner that describes how an event was performed.
- *IS The Instrument Of Event (Object, Event)*: Represent the information about the instruments that were involved in performing a certain event.
- *IS The Cause Of Event(Event, Event)*: Represent the cause-effect relation between events.
- *IS The Feeling Of Character, IS The Belief Of Character, IS The Intention Of Character*: Represent emotions and mental states of story characters.
- *IS The Property Of (Property, Thing)*: A general relation that can be used to represent properties about different concepts.

Figure 6 next presents the same taxonomy of the ontology from Protégé's perspective. Using this taxonomy, the designed knowledge model was able to represent the knowledge that was required to achieve the defined objectives i.e. answer the key(5W1H) questions about a story.

c. Composite and High Order Concepts

In the previous section, we presented the set of atomic and explicit concepts and relations that were identified and defined during the ontology designing process. These atomic concepts and relations were generic and recurrent in every story of corpus.

On the other hand, children stories incorporate other types of knowledge that is necessary to fulfill the meaning and semantics of the story; these are the composite or higher order concepts and relations as well as the implicit ones. Composite concepts are the ones that aggregates other concepts and relations within their definition i.e. conjunctions, disjunctions, negations of other concepts or relations. For example, the concept *BadEvent*is defined as a conjunction or union between the two classes of *Event* and *Bad*:

$$BadEvent \equiv Event \sqcap Bad$$

In addition, the composite concept *Bad Event* is also an implicit type of knowledge that was not explicitly mentioned in text, yet it can be inferred based on the atomic concepts that were explicitly mentioned in the text as well as the previous knowledge that is incorporated within the supporting ontology i.e. the T-Box assertion.

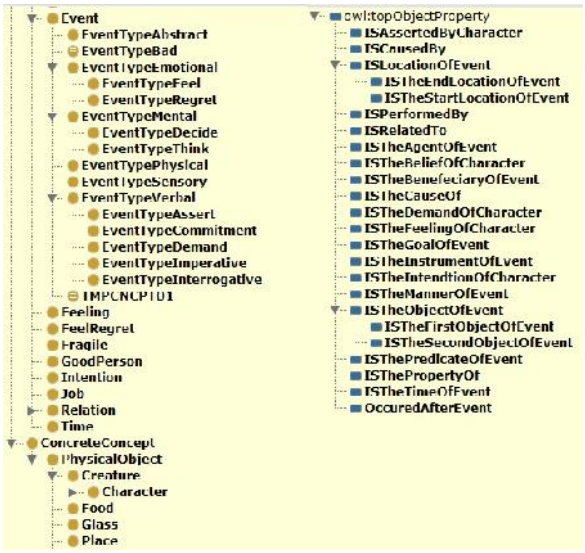

Figure 6: OWL Supporting Ontology, Classes and Object Relations

Similar to the designing process of atomic concepts and relations, it was necessary to opt for generalization when defining the composite and implicit oncepts and relations. This was important to ensure that the generated ontology is portable and sharable between the different stories. In addition, the generalization requirement complies with the initial findings of the corpus analysis process that were presented in section 4 . These initial findings concluded that the stories in our selected corpus were discussing actions that were performed by characters as well as the consequences of those events on story characters and the general theme or moral of the story.

For the purpose of this study, we narrowed the scope of the composite concepts and relations toones that can assist in informing whether a given character has performed or caused an event that resulted in pleasant or regrettable consequences. In addition, the defined (T-Box) assertions for the composite and the implicit concepts and relations should be able to infer if the character is good mannered or the opposite as well as the certain chain of events that are regrettable or not. Table 4 next presents some of the general class axioms that were defined in the ontology to satisfy the targeted

composite and implicit concepts and relations. The table presents the seaxioms using DL syntax, Manchester syntax as well as (ROWL) format (Sarker et al. 2016)Which might be more intuitive to users.

Table 4: Composite and Implicit Knowledge in the Supporting OWLontology

Concept	DL Syntax	Manchester OWLsyntax	ROWL Syntax
Event Type Bad	Event Type Bad ≡ Bad ⊓ Event	Bad and Event	Event(?e) and Bad(?e) → Event Type Bad(?e)
Did Some Bad Event	Did Some Bad Event ≡ Character ⊓∃is the Agent Of Event. Event Type Bad	Character and IS The Agent Of Event some Event Type Bad Sub Class Of Did Some Bad Event	Event Type Bad(?e) and Character(?a) and IS The Agent Of Event(?e, ?a) → Did Some Bad Event (?a)
Feel Regret	Feel Regret ≡ Character ⊓∃IS The Agent Of Event. Event Type Regret	Character and IS The Agent Of Event some Event Type Regret Sub Class Of Feel Regret	Event Type Regret(?e) and IS the Agent Of Event (?e, ?a) → Feel Regret(?a)
Good Person	Good Person ≡ Do Bad Event ⊓ Feel Regret	Do Bad Event and Feel Regret Sub Class Of Good Person	Do Bad Event (?a) and Feel Regret(?a) → GoodPerson(?a)

The defined composite and implicit concepts and relations are a subset of the knowledge types that can be established based on the extracted explicit knowledge. The purpose of this succinct definition was to assist in assessing the potentials of using DL in such scenarios while the alternatives for extending the defined composite concepts and relations can be covered in a future work.

On the other hand, Description Logic has some limitations in terms of representing assertions that have two variables in the second part of the proposition i.e. the consequent part. For that purpose, we employed Semantic Web Rules Language (SWRL) to deliver the required types of assertions. SWRL was selected because of its convenience as well as its support within Protégé that seamlessly incorporate SWRL rules within the reasoning process. Table 5 below demonstrates two SWRL rules that were defined to infer causal relations between two related events or chain of events.

Table 5: Supporting SWRL Rules

SWRL Rules
Character(?c) ^ IS The Cause Of(?e1, ?e2) ^ Event(?e1) ^ Event(?e2) ^ IS Performed By(?e1, ?c) -> IS Performed By(?e2, ?c)
IS The Cause Of(?e1, ?e2) ^ Character(?c) ^ Event(?e1) ^ Event(?e2) ^ IS The Agent Of Event(?c, ?e1) -> IS the Agent Of Event(?c, ?e2)

These rules indicate that the agent of a given event e.g. *Event1*, is the agent of a second event e.g. *Event2* if and only of *Event1* caused *Event2*.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we presented the implementation of the information extraction model that was initially presented in (Fasha et al. 2017). In that work, the authors presented a high-level description for a proposed model that can extract information from more relaxed versions of Arabic text domains. In addition, that work presented a work methodology that can be followed to implement the proposed information extraction model. In this work, we examined the initial steps of the proposed work methodology which include the selection of a suitable corpus, the analysis of the corpus and annotation it using a custom POS tagging scheme. We also discussed the method that we employed to design the supporting knowledge

representation model that shall be used to represent the extracted information. In addition, the initial preparation steps were enabled by a preparing Java-based modules that can facilitate the implementation of the corpus annotation process as well as the designing of the support OWL-based ontology.

The efforts presented in this work set the basis for our future work which includes implementing the other steps of the proposed model in (Fasha et al. 2017), which includes the designing of the two-phases information extraction process based on the established POS annotations and the supporting OWL ontology.

REFERENCES

1. *Al-Thubaity AM, Khan M, Alotaibi S, Alonazi B (2014) Automatic Arabic term extraction from special domain corpora. Proc Int Conf Asian Lang Process 2014, IALP 2014 1–5 . doi: 10.1109/IALP.2014.6973468*
2. *Al Zamil MGH, Al-Radaideh Q (2014) Automatic extraction of ontological relations from Arabic text. J King Saud Univ - Comput Inf Sci 26:462–472 . doi: 10.1016/j.jksuci.2014.06.007*
3. *Alruily M, Alghamdi M (2015) Extracting information of future events from Arabic newspapers: An overview. Proc 2015 IEEE 9th Int Conf Semant Comput IEEE ICSC 2015 444–447 . doi: 10.1109/ICOSC.2015.7050848*
4. *Attia M (2006) An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite* State Networks. Challenges Arab NLP/MT Conf … 48–*67*
5. *Bentrcia R, Zidat S, Marir F (2017) Extracting semantic relations from the Quranic Arabic based on Arabic conjunctive patterns. J King Saud Univ - Comput Inf Sci. doi: 10.1016/j.jksuci.2017.09.004*
6. *Bird S (2006) Nltk. Proc COLING/ACL Interact Present Sess - 69–72 . doi: 10.3115/1225403.1225421*
7. *Boudlal A, Lakhouaja A, Mazroui A, et al (2010) Alkhalil Morpho SYS1: A Morphosyntactic Analysis System for Arabic Texts. Int Arab Conf Inf Technol 1–6*
8. *Buckwalter T (2002) Arabic Morphological Analyser Version 1.0. Linguist Data Consort numéro LDC2002L49 2002*
9. *El-Salam MA, Hussein AK, Fahmy AA (2013) Understanding a simple Arabic stories using event calculus. Am J Appl Sci 10:1298–1306 . doi: 10.3844/ajassp.2013.1298.1306*
10. *Fasha M, Obeid N, Hammo B (2017) A Proposed Model for Extracting Information from Arabic-Based Controlled Text Domains*
11. *Kim ZM, Jeong Y, Choi H (2016) Understanding News Stories through SVO Triplets. 498–501*
12. *Madar C, Sarmistha L (1989) A Story Understander using Rhetorical Structures. Ieee 54–57*
13. *Manning C, Surdeanu M, Bauer J, et al (2014) The Stanford CoreNLP Natural Language Processing Toolkit. Proc 52nd Annu Meet Assoc Comput Linguist Syst Demonstr 55–60 . doi: 10.3115/v1/P14-5010*
14. *Mesmia F Ben, Zid F, Haddar K, Maurel D (2017) ASRextractor*: A Tool extracting Semantic Relations between *Arabic Named Entities*

15. *Azzoug Omar, Learning English and Arabic Two different Methodologies: Case of Modern Standard Arabic, International Journal of English and Literature (IJEL), Volume 5, Issue 4, July-August 2015, pp. 123-132*
16. *Omri H, Neji Z, Ellouze M, Hadrich Belguith L (2017) The role of temporal inferences in understanding Arabic text. Procedia Comput Sci 112:195–204 . doi: 10.1016/j.procs.2017.08.228*
17. *Pasha A, Al-badrashiny M, Diab M, et al (2014) MADAMIRA*: A Fast , Comprehensive Tool for Morphological *Analysis and Disambiguation of Arabic. Proc 9th Lang Resour Eval Conf 1094–1101*
18. *Riloff E, Riloff E (1999) Information extraction as a stepping stone toward story understanding. Underst Lang Underst Comput Model Read 435–460 . doi: 10.3115/1119176.1119205*
19. *Sadek J, Meziane F (2016a) Extracting Arabic Causal Relations Using Linguistic Patterns. ACM Trans Asian Low-Resource Lang Inf Process 15:1–20 . doi: 10.1145/2800786*
20. *Sadek J, Meziane F (2016b) A Discourse-Based Approach for Arabic Question Answering. ACM Trans Asian Low-Resource Lang Inf Process 16:1–18 . doi: 10.1145/2988238*
21. Samson BP, Ong E (2014) Extracting Conceptual Relations from Children's Stories. Lncs 8*863:195–208*
22. *Sarker MK, Carral D, Krisnadhi AA, Hitzler P (2016) Modeling OWL with rules: The ROWL protégé plugin. CEUR Workshop Proc 1690:1–4*
23. *Sawalha M, Atwell E, Abushariah M a. M (2013) SALMA Standard Arabic Language Morphological Analysis. In: Proc. ICCSPA Int. Conf. Commun. Signal Process. their Appl. Sharjah, UAE*
24. *Seyoufi H (2014) My Morals and Behaviour Series. Kitabi Institution for Publishing and Distribution*
25. *Souteh Y, Bouzoubaa K (2015) SAFAR platform and its morphological layer SAFAR platform and its morphological layer*
26. *Uschold M, King M (1995) Towards a Methodology for Building Ontologies. Methodology 80:275–280 . doi: 10.1.1.55.5357*
27. *W3C OWL Working Group (2012) OWL 2 Web Ontology Language. W3C Recomm 1–16*
28. *Wang W, Zhao D (2012) Ontology-based event modeling for semantic understanding of Chinese news story. Commun Comput Inf Sci 333 CCIS:58–68 . doi: 10.1007/978-3-642-34456-5_6*
29. *Zhou X, Wu Z, Yin A, et al (2004) Ontology development for unified traditional Chinese medical language system. Artif Intell Med 32:15–27 . doi: 10.1016/j.artmed.2004.01.014*