ACIT'2017
The International Arab Conference on Information Technology
Yassmine Hammamet, Tunisia
December 22-24, 2017

Towards Building a Frame-Based Ontology for the Arabic Language

Mariam Biltawi, Arafat Awajan and Sara Tedmori
Princess Sumaya University for Technology, Jordan

***Abstract*— This paper proposes a framework for building a frame-based ontology for the Arabic language. The proposed framework consists of two main phases. The first phase involves manual construction of a seed frame-based ontology. This is followed by the second phase in which the seed frame-based ontology is enriched with new lexical fields (only if they do exist), and/or enriching it with binary relations between existing or new lexical fields. The binary relations considered are synonyms/antonyms, hyponyms/hypernyms, and holonyms/meronyms. In addition, the paper presents a comprehensive introduction of lexical semantics providing examples from the Arabic language, and then surveys works of researchers aiming to build ontologies for the Arabic language.**

Keywords—Arabic ontology; lexical semantics; natural language processing

I. INTRODUCTION

Challenges in Natural Language Processing (NLP) frequently involve, amongst many others, the subtask of Natural Language Understanding (NLU). After identifying the syntactic structure of text, NLU focuses on analyzing the semantic features present in this text such as concepts, entities, keywords, relations, emotions, categories and many more, with the anticipation of gaining full understanding of the meaning conveyed in the text. Semantics refers to the study of the meanings of words and phrases in language and can be applied to single words (aka lexical semantics) or to entire texts (aka compositional semantics). Lexical semantics is concerned with the meanings of individual words and with the meaning/semantic relationships that individual words have with one another; in addition to the semantic features that help distinguish similar words. Compositional semantics is concerned with the meaning of the sentence or larger unit which goes beyond simply combining the meaning of the individual lexical words/units.

One important aspect of lexical semantics is finding out how individual words/units relate to one another. Lexical relationships are usually used to indicate both semantic and associative relationships among words, including phonetic, morphological, and morpho-syntactic relations. For example, in relation to words having multiple meanings (senses), a polysemous word is one word with different meaning; whilst homonymous words are essentially different words that have the same spelling and pronunciation but different meanings [1]. Some lexical relationships are symmetric such as synonyms/antonyms, while other lexical relationships are hierarchal such as hyponyms/hypernyms and holonyms/meronyms. Examples of additional lexical relations include entailment (e.g.: sad / حزن vs. cry / بكاء), magnifier (e.g.: wound / جرح vs. badly / بشكل سيء), singular/plural (e.g.: (sheep / خروف vs. flock / قطيع), idiom and operator (question / سؤال vs. ask / يطلب). Figure 1 illustrates the semantic relationships in a hierarchal form.

Synonyms refers to words that have different pronunciation but share the same meaning (e.g sit / جلس – قعد); whereas antonymy refers to word pairs sharing opposite meanings (e.g hot and cold / ساخن وبارد). Hyponymy refers to the relationship between a general word (aka hypernyms) and specific instances of it (aka hyponyms). For example, (cats and dogs / القطط و الكلاب) are hyponyms of the hypernym word (animals / حيوانات). Hyponymy/hypernymy are considered asymmetric relationships. Another hierarchal relation is holonym/meronym. Holonym refers to a word that denotes a whole of another word namely meronym which in turn is a part of the holonym. For example: the meronym (عجل / wheel) is a part of the holonym (سيارة / car).

Homonymy refers to words that have different meanings but share the same spelling, whereas homophones refers to words that also have different meanings but are pronounced the same. Examples of homophones and homographs exist in the English language, while only examples of homographs can be found in the Arabic language as Arabic language is highly phonetic, i.e.: the writing reflects the pronunciation. For example, the word (ذهب) is a homograph word because it has two meanings either 'went' or 'gold' and one spelling. Example on homophones for English language; (Two, to and too), (Flour and flower). WordNet [2] is an example of a resource that provides binary lexical relations between words. Arabic WordNet [3], a WordNet for Arabic language, is also available.

Semantic fields is a set of words that have related meaning to specific object or namely a frame. These semantic fields represent n-ary relations with the frame that they refer to through capturing more relationships among entire sets of words from a single domain. For example; the words: University / جامعه, Lecturer / محاضر, Student / طالب, Hall / قاعة, Library / مكتبة, Lab / مختبر, Section / شعبة, Course / مادة, Registration / تسجيل, are all related to the frame University Education / تعليم جامعي. The realization of lexical semantics is possible after successfully completing some/all of following subtasks: word sense disambiguation, semantic role labeling, multiword expression composition/decomposition, ontology

learning and population, and semantic language modelling. These lexical semantics of the n-ary relations can be used to define event structures, which are mainly in the form of predicate-argument arrangement. For example, in the event (ابراهيم كتب رسالة / Ibrahim wrote a letter) the predicate here is (كتب / wrote) and the arguments are (ابراهيم ، رساله / letter, Ibrahim) which represent the subject and object respectively. In this case, the subject represents the semantic role agent (i.e. the writer Ibrahim) and the object represents the semantic role patient (i.e. what is written and in this example it is the letter), semantic roles are also called thematic roles of the arguments. Another observation is the semantic restriction, also called selectional restriction, and means that the word enforces on the environment in which it occurs, for example we cannot say (ابراهيم كتب الطاولة - Ibrahim wrote a table). Different types of thematic roles exist including agent, experiencer, theme, result, force, instrument, content, goal, source, beneficiary, and others. PropBank [4], NomBank [5], VerbNet [6], and FrameNet [7] are examples of freely available predicate models for the English language, while there is only two predicate models for the Arabic language; the Arabic PropBank [8] and the Arabic VerbNet [9].

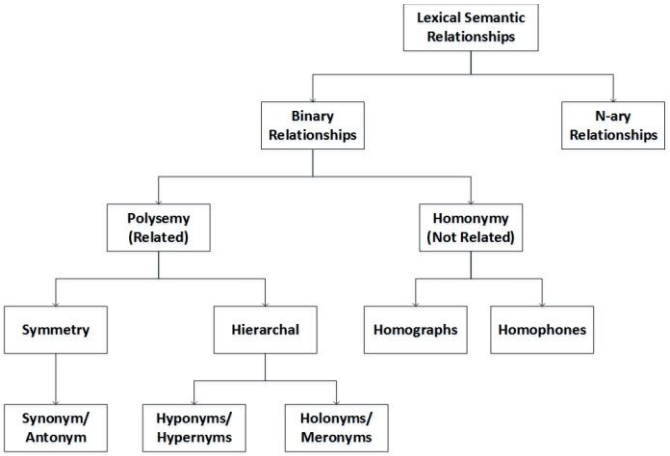

Fig. 1. Hierarchy of binary semantic relationships

The meaning of a given linguistic expression, irrespective of the language, can be uncovered through sophisticated linguistic analysis which attempts to find the correct semantic representation inherit in it and can be modeled through mapping the expression into three main elements; objects, properties of objects, and relations among objects, using meaning representation schemes such as predicate logic, description logics, semantic networks, and frames. For example, in the sentence mentioned previously (ابراهيم كتب رسالة / Ibrahim wrote a letter), the objects are (ابراهيم، رساله / letter, Ibrahim) which represents grammatically the subject (Ibrahim / ابراهيم) and the object (a letter / رسالة). The properties of these objects are derived from the predicate (wrote - كتب), implying that the subject "Ibrahim" is "the writer" while the object "a letter" is "what written". Next, the relations between objects, as previously stated Ibrahim represents a writer, he can also be listed under the category Person. Suppose the sentence was (Ibrahim wrote a poem) in this case the intersection between the objects, Ibrahim and poem, will give the relation "Ibrahim is a poet", which also represents a sub category of the writer category. Here comes the role of ontologies through capturing the superset/subset relations among objects. Figure 2 illustrates the superset/subset relations of the example. The term "ontology" comes from the field of philosophy that is concerned with the study of being or existence. In computer and information science, ontology is a technical term used to describe concepts, properties and relations among concepts in order to represent models of knowledge or discourse; thus, an ontology can represent meta-data schema for a knowledge of different applications in the form of vocabulary of concepts, and it can be shared by humans and machines [10].

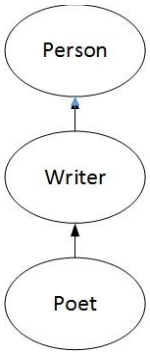

Fig. 2. Graphical representation of person ontology

Ontologies can also be used to integrate heterogeneous databases, enabling interoperability among disparate systems. Ontology learning is the process of automatic or semi-automatic construction of an ontology either from scratch, extending an existing one in order to enrich it, integrating existing ontologies, or adapting a generic ontology for a specific domain [11]. Ontology population, on the other hand, is the process augmenting an existing ontology with instances of concepts and relations [11]. Therefore, the target of ontology population is the extraction of ABox (instances and facts) knowledge according to specific ontologies, disambiguating extracted instances with respect to well-known Linked Open Data (LOD) knowledge bases; whereas the target of ontology learning is the extraction of TBox (classes and properties) knowledge. Noy and McGuinness [12] state five reasons why to use ontologies. The first reason is that ontologies allow the sharing of common understanding between people and machines. Another reason to use ontologies is that it facilitates reuse of domain information. Ontologies allow you to make explicit assumptions. Ontologies allow the distinction between domain and operational knowledge, and finally it facilitates the analysis of domain knowledge.

The subtasks of ontology learning are generally divided into eight layers renowned as the Ontology Learning Layer Cake (OLLC) [13]. The processing of subtasks should start from the lower layer then proceed through the higher layers

starting from term acquisition up to general axioms through the steps:

1. Term acquisition,
2. Synonyms,
3. Forming concepts,
4. Organizing concepts hierarchically,
5. Learning relations and properties within a domain,
6. Organizing relations hierarchically,
7. Instantiate axiom schemata and their definitions.

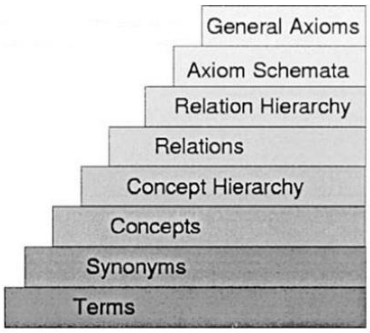

Fig. 3. Ontology Learning Layer Cake (OLLC) [13].

The goal of this paper is to present a framework for building a simple frame-based ontology, which starts by manual frame-based ontology construction and then is automatically enriched with lexical fields and binary relationships. The rest of this paper is organized as follows, section 2 surveys existing methods for building ontologies for the Arabic language, section 4 presents the proposed framework, and section 5 presents the conclusion.

II. Related Work

Ontologies can be built either from unstructured text, or through exploiting directories of web documents, or through integrating existing resources. In this section, the authors provide details of attempts made by researchers to build ontologies for the Arabic language. The majority of the research works focusing on building Arabic ontologies target only the first four layers of the OLLC. Al-Arfaj and Al-Salman [14] in their paper, provide a survey of Arabic NLP tools that can be used in the preprocessing phase when constructing an ontology. These tools are capable of manipulating Arabic text either by splitting sentences, or by facilitating tokenization, and/or part-of-speech tagging. Mazari et al. [15] proposed an approach for automatic building of domain ontology from Arabic text using a statistical technique. The authors collected a corpus of 57 documents comprising 468,554 words, from Arabic books and journal articles. Then the corpus was preprocessed by word segmentation and normalization, stop words deletion before and after light stemming, and light stemming. The processing phase comprised of extracting the repeated segments, and extracting the co-occupants. The resulted file contains two co-occurring terms along with their frequencies and the co-occurring frequency, which still remains to be validated by an expert.

Albukhitan and Helmy [16] proposed an ontology learning system for the Arabic language, which consists of seven steps; first the document format is analyzed to be prepared for the second step which is NLP processing where basic and advanced NLP tasks are applied. According to the authors, the basic NLP tasks are sentence splitting, phrase chunking, tokenization, token normalization, POS tagging, stemming, and co-references; while the advanced NLP tasks are parsing, discourse analysis, semantic role labeling, key phrase extraction, polarity analysis, and morphological analysis. The third step is concepts recognition using statistical and clustering methods. In step four taxonomic relations were recognized through pattern based algorithm, and concept hierarchies were driven. In step five, non-taxonomic relations were recognition. In step six, the ontology is constructed and finally it is written in some ontology language. For testing purposes, the authors developed a prototype of the proposed system, and manually annotated 100 documents.

Al-Arfaj and Al-Salman [17] proposed a framework for building an ontology from Al-Hadith corpus. The framework consists of four main steps, preprocessing, concept extraction, relation extraction, and ontology edition. The authors did not implement their work, but they provided a survey of available methods of building ontologies for the Arabic language along with the challenges faced when building ontologies for the Arabic language with suggested solutions. Another proposed an architecture for extracting ontology from Arabic corpus namely ArabOnto [18]. Al Zamil and Al-Radaideh [19] proposed a methodology to extract semantic relationships from Arabic text having certain patterns. Their approach is based on an enhanced version of Hearst's algorithm [20].

Belkredim and El-Sebai [21] discussed the ontological representation for the Arabic language, providing a design for their ontology which is based on the relations between the morphological classes of the Arabic language, mainly on verbs. No implementation was provided. Other research papers used association rules to extract ontologies from the Arabic text such as Quran corpus and Hadith corpus. Harrag et al. [22] proposed a combined approach of using pattern based schemes and association rules to extract Quran ontology through extracting concepts and semantic relation. In another research, association rules were implemented on Hadith [23].

Benaissa et al. [24] proposed a semi-automatic construction approach of lexical ontology from Arabic text. The approach is based on the synonymy relations between verbs and a clustering method used to exploit graphs of synonymy relations. The work is done using the Arabic dictionary (المعجم الغني) and the authors developed ontoArab-Maker tool to implement their approach. The evaluation is conducted manually on a sample of two verbs from the resulting ontology. Ishkewy et al. [25] proposed a lexical ontology namely Azhary for the Arabic language. Azhary consists of 26,195 Arabic words grouped into synonyms named synsets, thus their ontology contained 13,328 synsets. They also recorded other

binary relationships among these words. The building of their ontology is dependent on creating a seed of words containing 77,439 words from the Holy Quran.

Other attempts to build ontologies involve exploiting directories of web documents. For example, Al- Rajebah et al. [26] proposed an automatic approach for building ontology from Wikipedia. Semantic relations are extracted from Arabic Wikipedia articles and their approach contains two main components, the XML parser and the ontology generator. On the other hand, Halawani [27], proposed a framework for a multi-disciplinary ontology building from multiple resources for the Arabic language, his approach consisted of two main phases. In the first phase, the Arabic ontology is built from multiple resources through extracting categories and relations and preprocessing the textual contents. Then in the second phase, the ontology is enhanced and enriched.

Another way for building ontology is through integrating available resources such as predicate models, other existing ontologies, but for the Arabic language there is no such research, due to the lack of freely available resources. Although a number of attempts have been made to build ontologies for the Arabic language, only one resource, Arabic WordNet, which provides a lexical ontology for the Arabic language is freely available.

III. Proposed Framework

The main idea of the proposed approach is to provide machines with ontologies built from picture dictionaries such as "the new oxford picture dictionary", "the Heinle picture dictionary" and "the word by word picture dictionary", imitating the idea that children learn from such dictionaries in their early ages, and they are provided with the basic terms related to the world. This ontology is enriched with binary relations between the existing terms, and will be enriched with new terms that relate to the existing ones. Thus, the idea starts with frames having n-ary relation with semantic fields. And each semantic field will also represent a frame that has n-ary relation with other semantic fields forming a hierarchal structure. Then this hierarchal structure will be enriched with the binary relations creating a graph of terms that relate to each other.

The proposed ontology learning and population framework consists of two main phases. In phase one, a preliminary simple ontology is constructed manually through exploiting picture dictionaries. This results in a number of frames, where each frame contains a group of semantic fields, as illustrated in figure 4. In phase two, the WordNet is exploited to search for binary relationships between the semantic fields already prepared in the previous phase, along with enriching the existing ontology with other terms related to the existing fields.

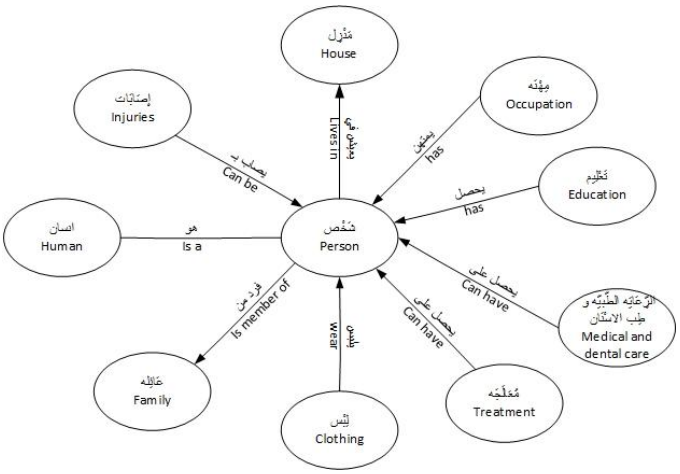

Fig. 4. Person Frame

A. *Phase 1: Manual Construction of the Frame-based Ontology.*

This phase represents the manual construction of frame-based ontology from the three picture dictionaries stated previously, the authors focused on using mainly "the new oxford picture dictionary", figure 5 illustrates a page from this dictionary that represents the frame "living room" and the lexical fields listed below the picture. This phase consists of the following manually conducted steps:

1. Data collection: 66 frames were collected from three English picture dictionaries and mainly from "the new oxford picture dictionary", each frame having multiple terms/semantic fields related to it.
2. Grouping: frames having relations with each other were grouped together to form one super frame, thus to facilitate the construction of a hierarchal representation. in step 4.
3. Removing, adding and translating into Arabic: these 66 frames with their related lexical fields will be translated and those fields that have unrelated meanings will be eliminated, while some other important fields will be added. For example, the word "part" was a lexical field under the frame "human head", which is considered not an important word and has no significant translation into Arabic related to the frame "human head". Another example, the word "cousin" in the frame "family" represents different meanings in the Arabic language, a female or a male cousin either from the father's side or from the mother's side, therefore all these meaning needed to be listed.
4. Hierarchal preparation: After the frames are grouped, cleaned, and translated, relations between these frames will be built manually in order to prepare the data in a hierarchal form which constitutes a frame and lexical fields below each frame, thus creating the frame-based ontology. For example, nine frames can represent "lexical

fields" under the frame "Person / شخص", noting that there is no such frame named "Person" in the picture dictionary, but we can group most of the frames that are either part of a person's life or a person is part of it together to form the super frame "Person". Figure 4 illustrates the nine frames/lexical fields that related to the main frame "Person / شخص". The relation in the figure are depicted using edges, the undirected edge indicates synonymy while the directed edge indicates hierarchal relation, for example the relations; "a person is a human", could represent synonymy. "A person has occupation" an occupation is a part of a person's life. Each one of these nine sub-frames has a number of lexical fields, which in turn may represent a sub-frame, forming a hierarchy. The sub-frame "family / عائلة" for example, have the lexical fields depicted in figure 6, these lexical fields of the frame "family" represent parts of the "family". Other frames can further have one or more sub-frames, as an example the lexical field "House / منزل" represents a sub-frame, related with the frame "Person" through the relation, "Person is part of House". House has the lexical fields depicted in figure 7, and each one of these lexical fields can also represent a sub-frame, for example the sub-frame "Living room" depicted in figure 8.

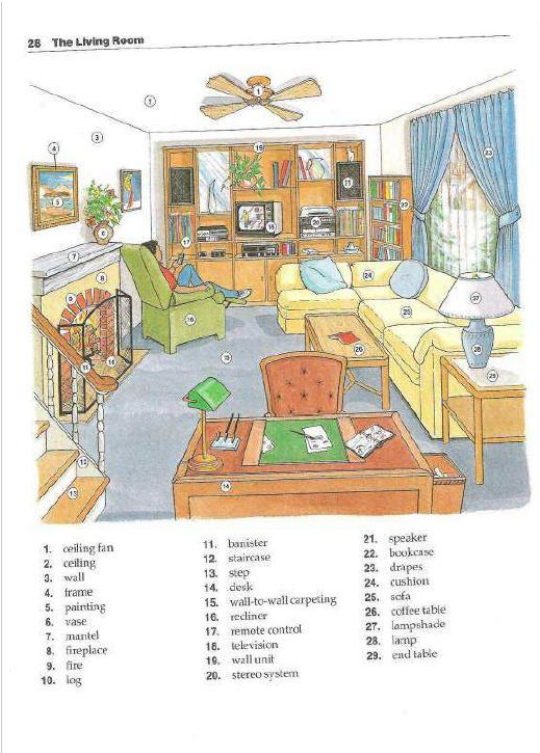

Fig. 5. Living Room Screenshot from "The New Oxford Picture Dictionary"

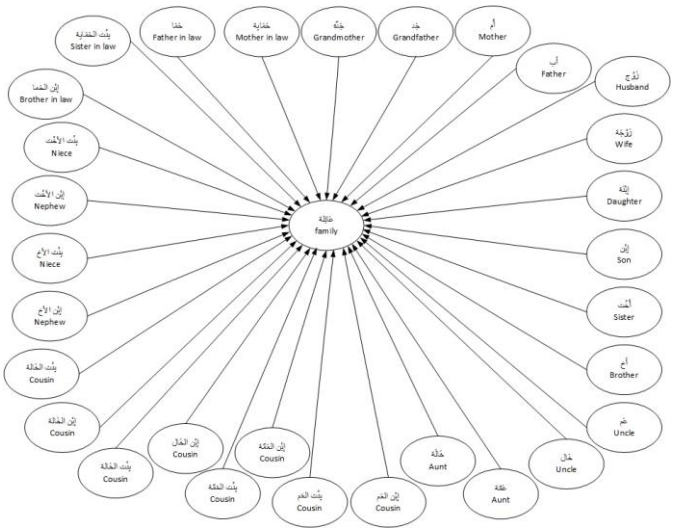

Fig. 6. Family

B. *Phase 2: Frame-based Ontology Population.*

This phase represents an automatic enrichment for the manually constructed frame-based ontology, created in phase one, in order to populate the ontology with new terms or relations through exploiting existing ontologies such as the WordNet. This phase consists of the main step: Term lookup and enrichment, which is applied through looking up for each frame of the frame-based ontology in the WordNet. If a match is found, then all its related terms and relations will be brought and the ontology is populated. The resulted ontology will represent a tree, where nodes represent frames and edges represents relationships. For example, for the frame "Person / شَخص" a number of senses will be brought from the WordNet. Such as (شَخْص ، شَخْص مَا، أَحَد ما، فَرْد، نَفسْ، رُوح، إنسان، اِمْرُؤ، مَرء), where all of them could represent synonyms except (نَفسْ، رُوح) which means soul and can represent a part of "person". Therefore, the frame "Person" will be modified according to the new data received, noting that the replicated words will be neglected such as the senses (Person / شَخْص, Human / إنسان), and the modified frame "Person" will be as shown in figure 9, the shaded ovals are the newly added fields.

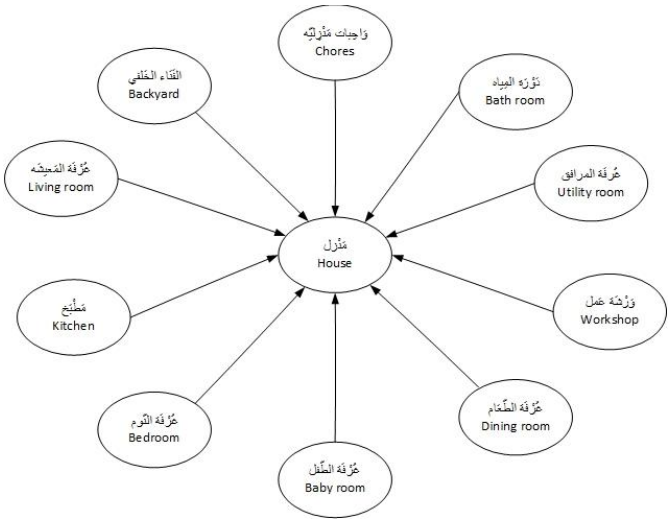

Fig. 7. Home

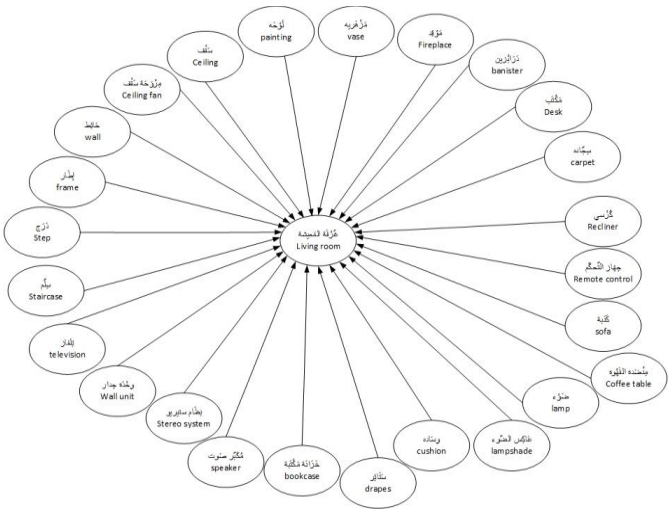

Fig. 8. Living Room

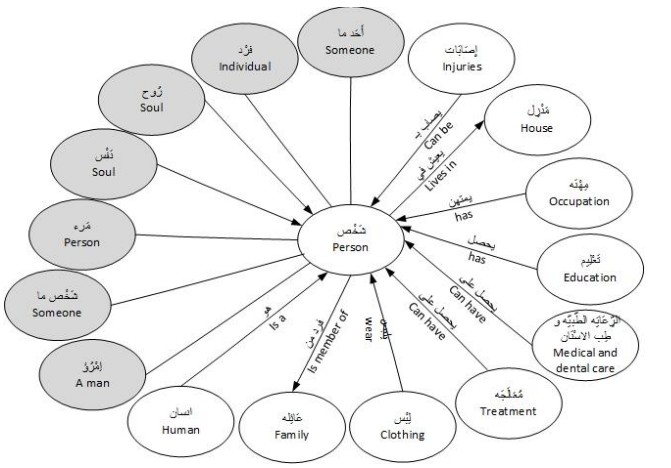

Fig. 9. Updated Frame of Person

IV. Conclusion

This paper proposed a framework for building Arabic frame-based ontology, which consisted of two main phases. The first phase concentrated on the manual construction of the preliminary frame-based ontology exploiting three picture dictionaries in the English language, an effort to remove, add, translate, and prepare the ontology in the hierarchical format is presented. On the other hand, the second phase concentrated on the automatic ontology population exploiting WordNet, in order to enrich the frame-based ontology with either new lexical fields, and/or with binary relationships between lexical fields. The resulted frame-based ontology represented a tree, where nodes are lexical fields and edges represented relationships. As a future work, an effort will be done to implement this work and conduct the experimental results.

References

[1] Jurafsky, Daniel, and James H. Martin. "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition."

[2] Miller, George A. "WordNet: a lexical database for English." Communications of the ACM 38, no. 11 (1995): 39-41.

[3] Black, William, Sabri Elkateb, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. "Introducing the Arabic wordnet project." In Proceedings of the third international WordNet conference, pp. 295-300. 2006.

[4] Kingsbury, Paul, and Martha Palmer. "From TreeBank to PropBank." In LREC, pp. 1989-1993. 2002.

[5] Meyers, Adam, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. "The NomBank project: An interim report." In HLT-NAACL 2004 workshop: Frontiers in corpus annotation, vol. 24, p. 31. 2004.

[6] Schuler, Karin Kipper. "VerbNet: A broad-coverage, comprehensive verb lexicon." (2005).

[7] Baker, Collin F., Charles J. Fillmore, and John B. Lowe. "The berkeley framenet project." In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1, pp. 86-90. Association for Computational Linguistics, 1998.

[8] Palmer, Martha, Olga Babko-Malaya, Ann Bies, Mona T. Diab, Mohamed Maamouri, Aous Mansouri, and Wajdi Zaghouani. "A Pilot Arabic Propbank." In LREC. 2008.

[9] Mousser, Jaouad. "A Large Coverage Verb Taxonomy for Arabic." In LREC. 2010.

[10] Drumond, Lucas, and Rosario Girardi. "A Survey of Ontology Learning Procedures." WONTO 427 (2008): 1-13

[11] Petasis, Georgios, Vangelis Karkaletsis, Georgios Paliouras, Anastasia Krithara, and Elias Zavitsanos. "Ontology population and enrichment: State of the art." In Knowledge-driven multimedia information extraction and ontology evolution, pp. 134-166. Springer-Verlag, 2011

[12] Noy, Natalya F., and Deborah L. McGuinness. "Ontology development 101: A guide to creating your first ontology." (2001).

[13] Cimiano, P.: Ontology Learning and Population from Text — Algorithms, Evaluation and Applications. Springer, Berlin–Heidelberg, Germany (2006) Originally published as PhD Thesis, 2006, Universitt Karlsruhe (TH), Karlsruhe, Germany

[14] Al-Arfaj, Abeer, and AbdulMalik Al-Salman. "Arabic NLP tools for ontology construction from Arabic text: An overview." In Electrical and

Information Technologies (ICEIT), 2015 International Conference on, pp. 246-251. IEEE, 2015.

[15] Mazari, Ahmed Cherif, Hassina Aliane, and Zaia Alimazighi. "Automatic Construction of Ontology from Arabic Texts." In ICWIT, pp. 193-202. 2012.

[16] Albukhitan, Saeed, and Tarek Helmy. "Arabic Ontology Learning from Un-structured Text." In Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on, pp. 492-496. IEEE, 2016.

[17] Al-Arfaj, Abeer, and Abdulmalik Al-Salman. "Towards ontology construction from Arabic texts-a proposed framework." In Computer and Information Technology (CIT), 2014 IEEE International Conference on, pp. 737-742. IEEE, 2014.

[18] Bounhas, Ibrahim, Bilel Elayeb, Fabrice Evrard, and Yahya Slimani. "ArabOnto: experimenting a new distributional approach for building Arabic ontological resources." International Journal of Metadata, Semantics and Ontologies 6, no. 2 (2011): 81-95.

[19] Al Zamil, Mohammed GH, and Qasem Al-Radaideh. "Automatic extraction of ontological relations from Arabic text." Journal of King Saud University-Computer and Information Sciences 26, no. 4 (2014): 462-472.

[20] Hearst, Marti A. "Automatic acquisition of hyponyms from large text corpora." In Proceedings of the 14th conference on Computational linguistics-Volume 2, pp. 539-545. Association for Computational Linguistics, 1992.

[21] Belkredim, Fatma Zohra, Ali El-Sebai, and Universitie Hassiba Ben Bouali. "An ontology based formalism for the arabic language using verbs and their derivatives." Communications of the IBIMA 11, no. 5 (2009): 44-52.

[22] Harrag, Fouzi, Abdullah Al-Nasser, Abdullah Al-Musnad, Rayan Al-Shaya, and Abdulmalik Al-Salman. "Using association rules for ontology extraction from a quran corpus." In Proc. 5th Int. Conf. Arabic Language Process., pp. 1-8. 2014.

[23] Harrag, F., A. Alothaim, A. Abanmy, F. Alomaigan, and S. Alsalehi. "Ontology extraction approach for prophetic narration (Hadith) using association rules." International Journal on Islamic Applications in Computer Science And Technology 1, no. 2 (2013): 48-57.

[24] Benaissa, Bedr-eddine, Djelloul Bouchiha, Amine Zouaoui, and Noureddine Doumi. "Building Arabic ontology from texts." (2015). Procedia Computer Science, vol: 73, pp: 7-15

[25] Ishkewy, Hossam, Hany Harb, and Hassan Farahat. "Azhary: An arabic lexical ontology." arXiv preprint arXiv:1411.1999 (2014).

[26] Al-Rajebah, Nora I., Hend S. Al-Khalifa, and AbdulMalik S. Al-Salman. "Exploiting Arabic Wikipedia for automatic ontology generation: A proposed approach." In Semantic Technology and Information Retrieval (STAIR), 2011 International Conference on, pp. 70-76. IEEE, 2011.

[27] Hawalah, Ahmad. "A Framework for Building an Arabic Multi-disciplinary Ontology from Multiple Resources." Cognitive Computation (2017): 1-9.