ISSN: 2321-7782 (Online)
Impact Factor: 6.047

Volume 4, Issue 8, August 2016

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Ontology Extraction from Text: Related works between Arabic and English languages

Dr. Mohamed Abd El-Fatah Belal[1]
Professor,
Computer Science Faculty of computers and information,
Helwan University
Cairo – Egypt

Dr. Hala Abdel-Galil[2]
Associate Professor,
Head of Computer Science Department
Faculty of computers and information, Helwan University
Cairo – Egypt

Yasser Mohamed Saber[3]
MSc student /Computer Science Department of
Faculty of computers and information, Helwan University
Cairo – Egypt

Abstract: Ontology base enables the sharing and reusing of information and allowing for the interoperation and integration of information systems. The automatic ontological relation extraction from the text is one of the important factors for representing the general documents and text in meaningful computerized way. The aim of this paper is to give a brief overview of ontology extraction approaches and levels that presented for the Arabic text and for the English text.

Keywords: Automatic extraction of Relationships, Arabic ontology; English ontology; ontology approaches; ontology extraction levels.

I. INTRODUCTION

Ontologies are used in artificial intelligence, the semantic web, systems engineering, software engineering, biomedical informatics, library science, enterprise bookmarking, and information architecture as a form of knowledge representation about the world or some part of it . The creation of domain ontologies is also fundamental to the definition and use of an enterprise architecture framework [5].Arabic language is currently the sixth most widely spoken language in the world. Itis the mother tongue of about 300 million of peoples. Arabic is an official language in more than 22 countries [7].So we need to highlight the ontology levels applied to the Arabic language with comparison with the English language.

II. ENGLISH BASED RELATED WORK

We have many types of research with different methods and levels introduced the ontology extraction from the text that applied to the English language. We will present here the introduced methods with some examples of the related researches. And what is the presented ontology level and what are the missing ontology levels for each method.

2.1 - Semi-automatic extraction method

This method is the starting approach that created in ontology extraction .It is basically depends on the human factor in defining the ontology roles and adjusting the extracted ontologies by user.

For example, this method is introduced for extracting binary as following:

- 2000-SEMI-AUTOMATIC ONTOLOGY-BASED KNOWLEDGE EXTRACTION [32], The study depends on calculating minimum link weight and maximum path length. the user is asked to identify the two concepts in the summary that are causally related and to identify the relationship between them. [32]

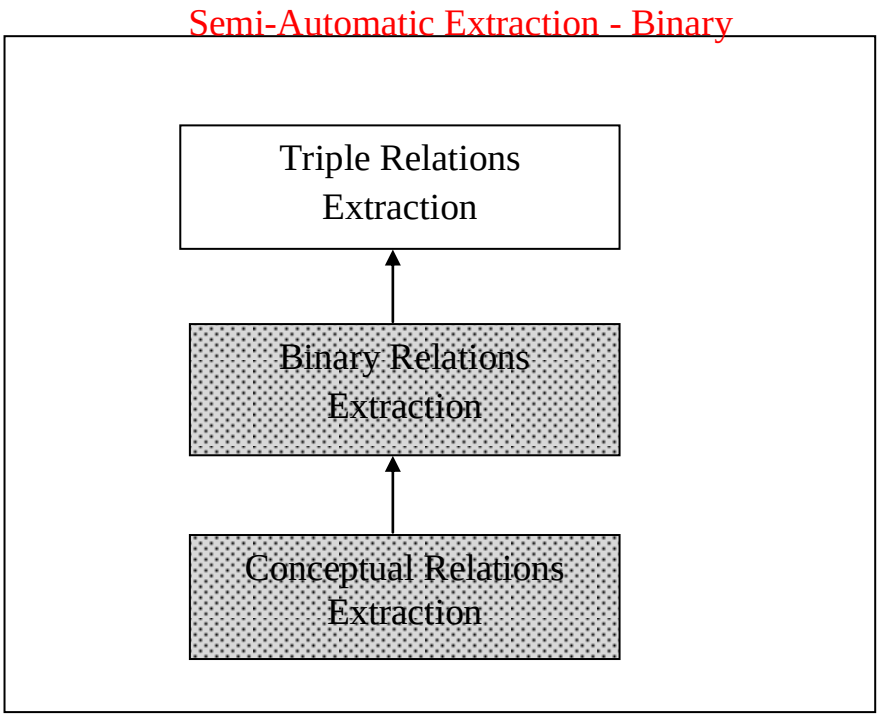

Figure 1: Semi-Automatic Ontology Extraction – English Language

- 2001-ONTOTERM knowledge base [31], the Semi-automatic approach that presents specific ontology on presorted data for a specific domain that enables the user to create new ontologies manually.
- 2002-TEXT-TO-ONTO framework [30] ,Using word Net and depends on applying machine learning techniques by calculating term frequencies in the text. This approach of semi-automatic ontology acquisition concedes extended support for manual ontology engineering.
- 2009-NIPF is presenting a semi-automatic development of an ontology library for the National Intelligence Priorities Framework (NIPF) topics. [29].It's more intelligent than the previous semi-automatic approaches as this approach implementing the defined binary relations automatically like (ISA, Part-Whole).

For example, This method is introduced for extracting triple as following:

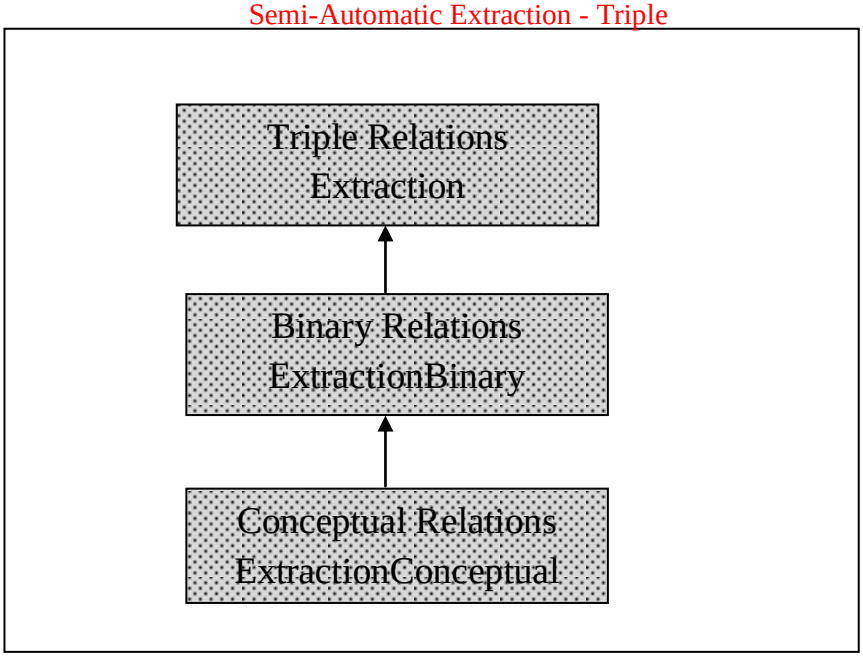

Figure 2: Semi-Automatic Ontology Extraction – English Language

2004-2005- ONTOLT, ONTOLT V1.0 framework [18,19,20], plug-in (OntoLT) for the widely used Protégé ontology development tool that supports the interactive extraction and/or extension of ontologies from the text. The OntoLT approach provides an environment for the integration

- of linguistic analysis in ontology engineering through the definition of mapping rules that map linguistic entities in annotated text collections to concept and attribute candidates (i.e. Protégé classes and slots).[19]

We have many tries from 2004 like our example as a good model for extracting the triple relation However it's regarded a semi-automatic extraction way.

The semi-automatic method that was introduced is presenting some binary and triple relations but it’s regarded as a weak method because simply no one or organization can define manually the whole domains or specific domain that must always have new relations. Also it does not present triple relations in many cases.

2.2 – Automatic binary relation extraction Level

Here we will introduce the automatic ontology extraction levels with the different approaches that the papers and studies applied to reaching it to present the ontology excretion result.

We will starting with the most reachable level, It is Automatic binary relation extraction Level. This Level is introduced in many kinds of methods as the following:

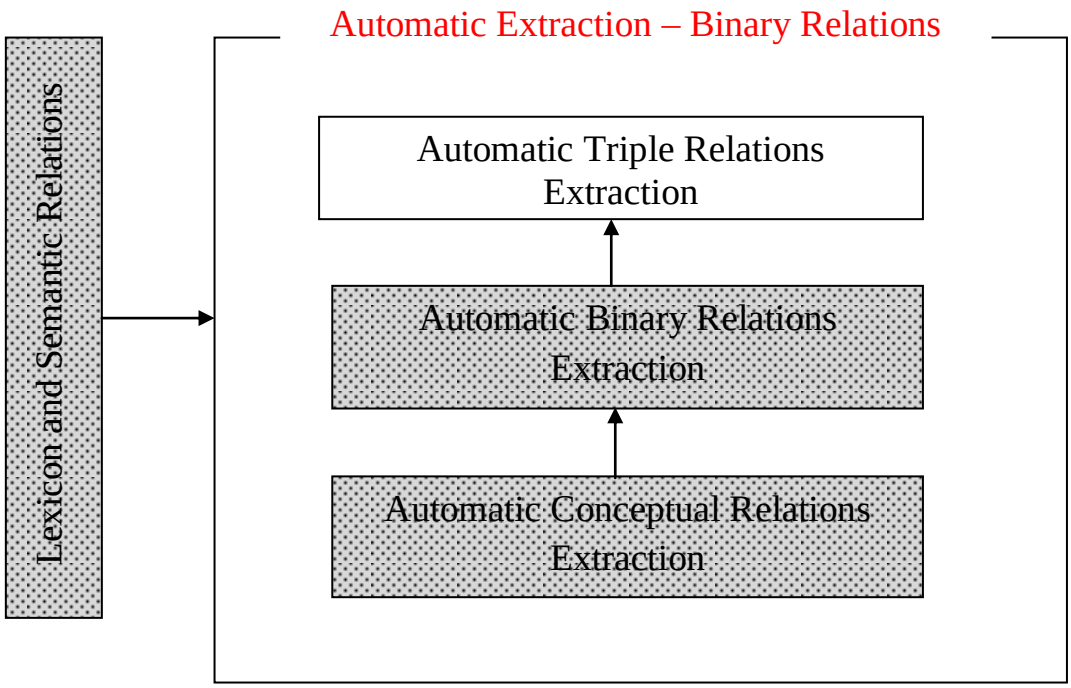

Figure 3: Automatic Ontology Relations Extraction – English Language

2.2.1 – Automatic binary Extraction from general text

For example, this method is introduced as following:

- 2010- Simple_Method_for_Ontology_Automatic_Extraction_from_Documents[22], It presents a method for automatic extraction of ontologies from documents or from a collections of documents that are independent of the document type and uses the junction of several theories and techniques, such as latent semantic for the extraction of initial concepts, WordNet and similarity to obtain the correlation between the concepts[22].

It's a good model for extracting the ontology concepts and Binary relation from general documents However the introduced way for extracting correlation between the concepts is limited by the defined binary relations like IS A , Part Of ,Subclass and equivalent relations So, It's not presenting the triple relations.

2.2.2– Query answering system binary extraction method

The studies that talking about Query answering systems focusing on the query parsing and matching between the query and the ontology extracted more than focusing on how is the extracting ontologies built. Also many of the introduced extraction ways ended by extracting binary relations and storing as a knowledge base.

For example , This method is introduced as following:

Keyword is a baseline system based on keyword matching. It directly matches the question substring containing the verb and the available argument with the input text, ignoring case and morphology. Given a match, two ways to derive the answer were considered: KW simply returns the rest of sentence on the other side of the verb, whereas KW-SYN is informed by syntax and extracts the answer from the subject or object of the verb, depending on the question (if the expected argument is absent, the sentence is ignored) .[23].

- **TextRunner** (Banko et al., 2007) is the state-ofthe-art system for open-domain information extraction. It inputs text and outputs relational triples in the form (R,A1,A2), where R is the relation string, and A1,A2 the argument strings. To answer questions, each triple-question pair is considered in turn by first matching their relation strings and then the available argument strings. If both matches, the remaining argument string in the triple returned as an answer. Results

were reported when exact match is used (TR-EXACT), or when the triple strings may contain the question ones assubstrings[23]

- **Resolver** (Yates and Etzioni, 2009) inputs TextRunner triples and collectively resolves coreferent relation and argument strings. To answer questions, the only difference from TextRunner is that a question string can match any string in its cluster. As in TextRunner, results were reported for both exact match (RS-EXACT) and substring (RS-SUB) . [23]
- **DIRT** (Lin and Pantel, 2001) resolves binary relations by inputting a dependency path that signifies the relation and returns a set of similar paths. To use DIRT in question answering, it was queried to obtain similar paths for the relation of the question, which were then used to match sentences[23].

2.2.3–Corpora based binary extraction method

Corpora is The plural form of a corpus .Corpus is a large collection of texts, It is a body of written or spoken material upon which a linguistic analysis is based[33].

The extraction here not from a general text , It's extracted from Corpora that collected and formed by a specific way like a list of words or list of sentences.The way of ontology extraction from corpora is often a statistical way that depends on calculating frequency of the couple words.

For example , This method is introduced as following:

- 2009-AUTOMATIC EXTRACTION OF LOGICALLY CONSISTENT ONTOLOGIES FROM TEXT CORPORA[34] , It's based on statistical extraction method and extracting from corpora text as well it's generating very good binary relations with semantic relations and using semantic software like WordNet.

The Corpora based extraction method that introduced is presenting binary relations that extracted from the corpus that is collected and formed by a specific way like a list of words or list of sentences.

Binary relation level can give us the relation between two words or between to concepts by defined classes like (IS A , Part Of , Sub of) but it can’t give us the meaning of the sentence as no presenting for the triple level.

2.3– Automatic triple relation extraction Level

Now we will continue introducing the automatic ontology extraction levels with the different approaches that the papers and studies applied to reaching it to presenting the ontology extraction result.

We will present the latest level that can be reachable till now; it is Automatic triple relation extraction Level.

This Level is introduced in many kinds of methods as the following:

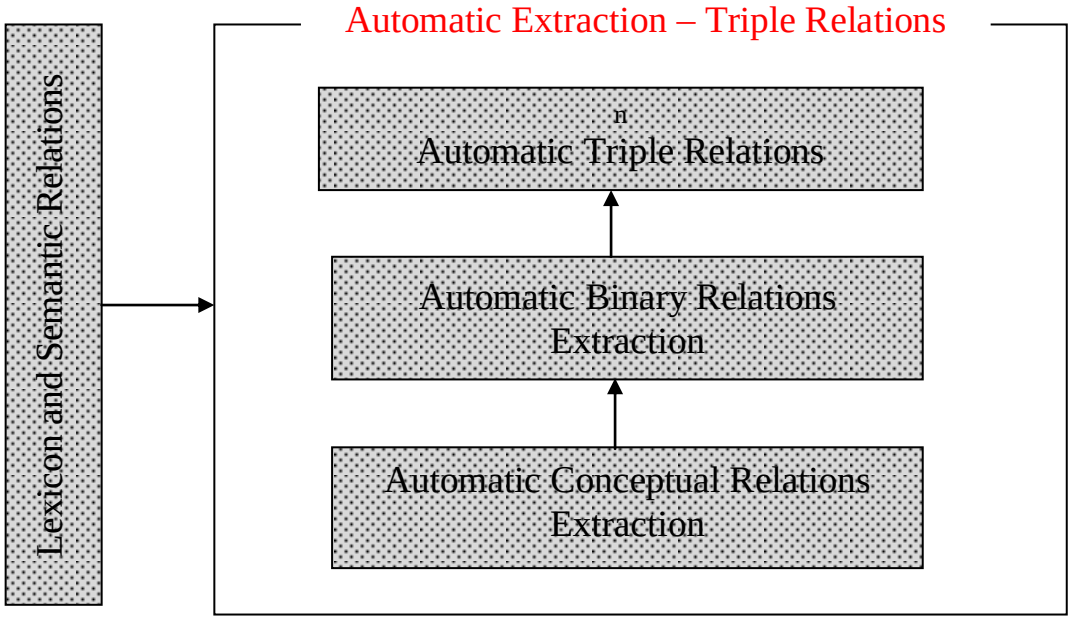

Figure 4: Automatic Ontology Relations Extraction – English Language

2.3.1 – Automatic triple Extraction from general text

For example, this method is introduced as following:

- 2013-Ontology Guided Information Extraction from Unstructured Text[17]. They using a triple extraction algorithm described in [Deli] that they have implemented using the

  StanfordCoreNLP java library. They extracting the subject, predicate and object.

- 2014-Concept relation extraction using Naïve Bayes classifier for ontology-based question answering systems[16] , Approach used in this research covering the most of the levels presented by us in the ontology extraction .

  The above two examples are extracting the triple relations automatically from general text using the Stanford parser and extracting the similarities using the WordNet or by identifying manually for the domain.

2.3.2 – Specific scheme triple extraction method

This is a development scheme method that developed specially to match specific formal data representation like ontology extraction from Wikipedia or extracting from documents organized as predefined scheme.

This type can't be applied to other documents or sources, except the source that designed especially for.

For example , This method is introduced as following:

- 2003-2005- Artequakt project [26 , 27 , 28] , a system which uses natural language tools to automatically extract knowledge about artists from multiple documents based on a predefined ontology [28] .

  The Artequakt presenting triple relation while The scheme building for Artequakt can be applied only to the artists' information.

- 2008- Okinet [25] Is a framework for Automatic Extraction of a Medical Ontology From Wikipedia . The Okinet presenting triple relation while The scheme building for Okinet can working only with the Wikipedia structure.

- 2013- Entity Extraction [24] system that automatically extracts RDF triples describing entity relations and properties from unstructured text. It's more intelligent than the previous as it can extract triple relations in RDF scheme while The scheme building for it seems that only can be working with the Wikipedia structure.

Specific scheme extraction method that introduced is presenting some triple relations But it's The Built in scheme working only on a specific data source and based on another defined scheme .So, It's can't represent general documents or whole specific domain.

2.3.3 – Query answering system triple extraction method

As we discussed the query answering system binary extraction, We have some examples for triple extraction. But we still notice that The studies that talking about Query answering systems focusing on the query parsing and matching between the query and the ontology extracted more than focusing on how is the extracting ontologies built.

For example , This method is introduced as following:

- **USP** (Poon and Domingos, 2009) parses the input text using the Stanford dependency parser learns an MLN for semantic parsing from the dependency trees, and outputs this MLN and the MAP semantic parses of the input sentences These MAP parses formed the knowledge base (KB). To answer questions, USP first parses the questions (with the question slot replaced by a dummy word) and then matches the question parse to parses in the KB by testing subsumption .[23]
- **OntoUSP** (Hoifung Poon and Pedro Domingos, 2010) uses a similar procedure as USP for extracting knowledge and answering questions, except for two changes. First, USP's learning and parsing algorithms are replaced with OntoUSP Learn and OntoUSP-Parse, respectively. Second when OntoUSP matches a question to its KB, it not only considers the lambda-form cluster of the question relation, but also all its sub-clusters[23].

III. Arabia Based Related Work

At variance with the English text domain, We have few number of studies with deferent methods and levels introduced the ontology extraction from the text that applied to the Arabic language. We will present here the introduced methods with some examples of the related researches .And what is the presented ontology level and what is the missing ontology levels for each method.

3.1 –Ontology terms extraction

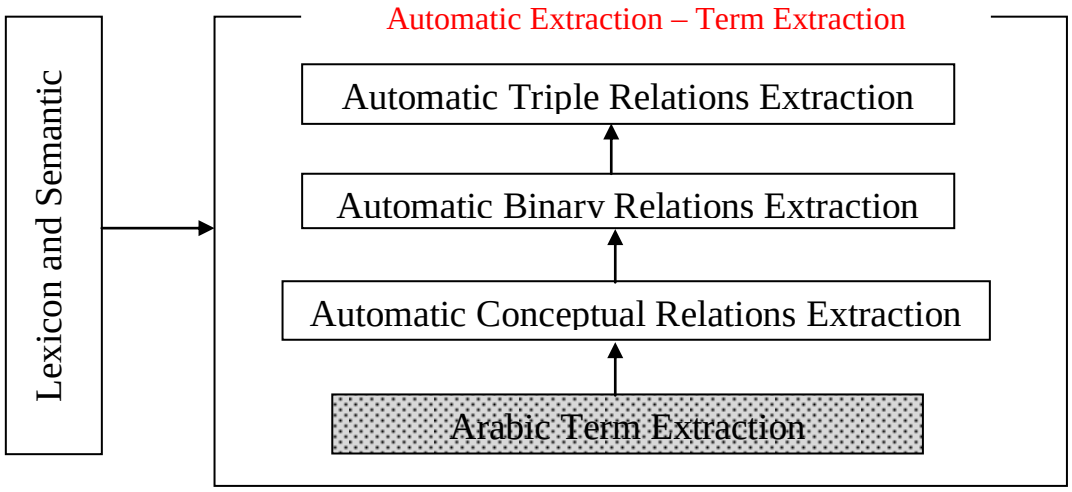

Figure 5: Automatic Ontology Relations Extraction – Arabic Language

As per the Arabic language have different needs, the approach that is introduced here is very important for building the Arabic ontologies. The term in Arabic language may be consist one word or more than one word together.

For example, This level is introduced as following:

2013-ARABIC TERM EXTRACTION USING COMBINED APPROACH ON ISLAMIC DOCUMENT [15] , approach used, frequency word-Reverse frequency document (TF.IDF)to calculate the frequencies each word in every document

Arabic terms extraction is a good step to starting the Arabic ontology extraction However we don't have here any type of relation between the words extracted like binary or triple relations.

3. 2 –Automatic Binary relation extraction Level

We will introduce the automatic ontology extraction levels that are introduced in the Arabic language domain to find relations between more that one word and more than one term.

Will presenting here the most reachable level for Arabic domain, it is Automatic binary relation extraction Level.

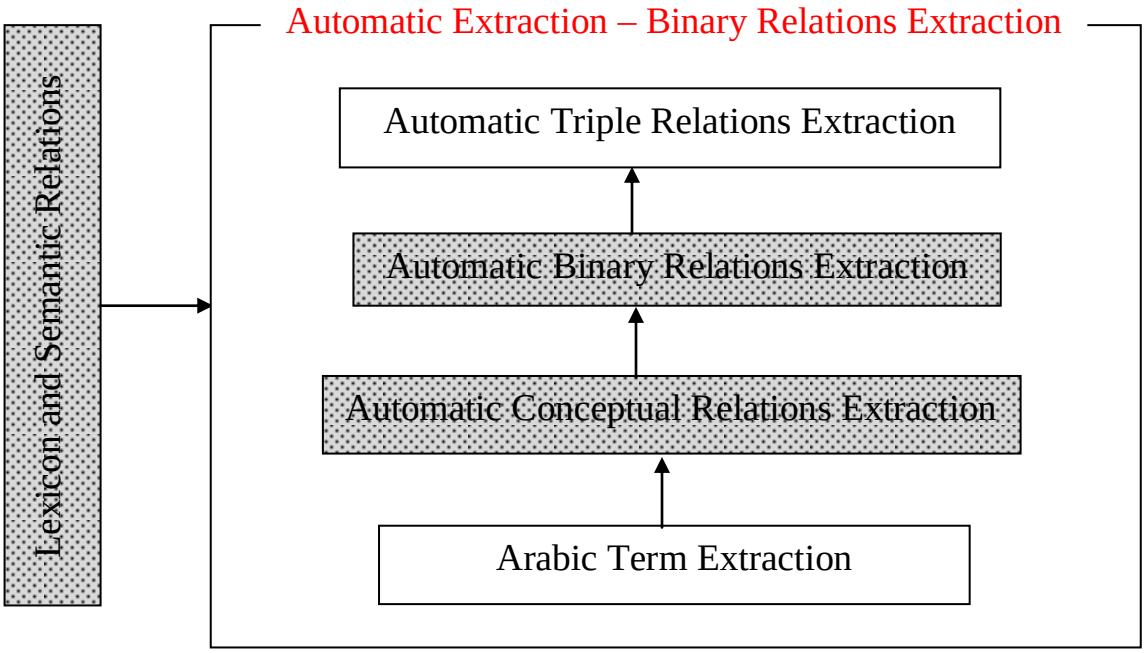

Figure 6: Automatic Ontology Relations Extraction – Arabic Language

This Level is introduced in a few studies as the following examples:

- 2014- Automatic extraction of ontological relations from Arabic text [14] , This research makes a good effort in Arabic binary relation extraction as well as using the Root extraction and concept extraction. It's presenting also the binary relation between the two words by using the defined classes like (isa هو وهي, Part of عضو في – من مكونات, has a يمتلك – له, kind of نوع من و احد الانواع) and adding some specially defined relation for the Arabic language like cause(بسيب).
- 2014- Azhary: An Arabic Lexical Ontology[13], The framework is presenting the ontology extraction to many binary relations and comparing the result with the Arabic WordNet(AWN). AWN is a famous Arabic lexical framework .

It presenting relations like (POSنوع الكلمة, synonym المعني,antonym المضاد, hypernymالأصل , hyponymالفرع ,holonym تحتوي علي, meronym توجد في)

- 2015- Ontology-Based Semantic Annotation of Arabic Language Text [12] ,It's also here Identifying the concepts and presenting the defined binary relation between them .

All above examples introduce Binary relation level that can give us the relation between two Arabic words or between to Arabic concepts by defined classes like (IS A , Part Of ,Sub of) but it can't give us the meaning of the sentence as no presenting for the triple level.

3.3–Triple in a query answering methods

To our knowledge, the QA studies that are introduced for the Arabic text domain not depending on ontological base for information extraction from the text .But it just depending on direct matching or semantic matching with the stored documents without extracting the ontologies from the text.

While we have case here that introduced a query processing and matching ontological answering to be Arabic interface for one of ontological query language (SPARQL) supposing that there is an ontology base already built and represented by this query language :

- 2015- An Ontology-Based Arabic Question Answering System[7], The core of the system is the approach we propose to translate Arabic NL queries to SPARQL. The approach makes intensive use of the ontology semantics to translate the user query to RDF triple patterns and infer any missing components to build up a complete SPARQL query. The proposed approach can process queries of different complexities and structures.[7]

So here triple relation extracted for the query only to matching the supposed ready ontology base represented by the concerned query language .And this not regarded an ontology extraction case.

3.4– Automatic triple relation extraction Level

To our knowledge and after many searching, No studies introduced the extraction triple relations from Arabic text till now. Research in Arabic domain introduced term extraction , concept extraction and Binary extraction .

And as mentioned in previous section 3.3 that few pieces of research built Arabic query to matching with ontological triple relation answering but by supposing that the ontology triples already extracted and represented in an ontological query language like SPARQL.

The Same idea found in other researches made for other purposes and assuming that the Arabic triple relation already created and stored. We have another example in AMASAT tool [10], It generating RDF Annotation (border) for the extracted binary ontology to collecting the property of same ontology in one RDF statement supposing that the ontology triples already extracted and represented in RDF statements.So the previous examples not regarded an ontology extraction case.

IV. CONCLUSION

We can summarize the result in the following comparison that illustrates the Automatic extraction for the ontological levels between the Arabic text and the English text.

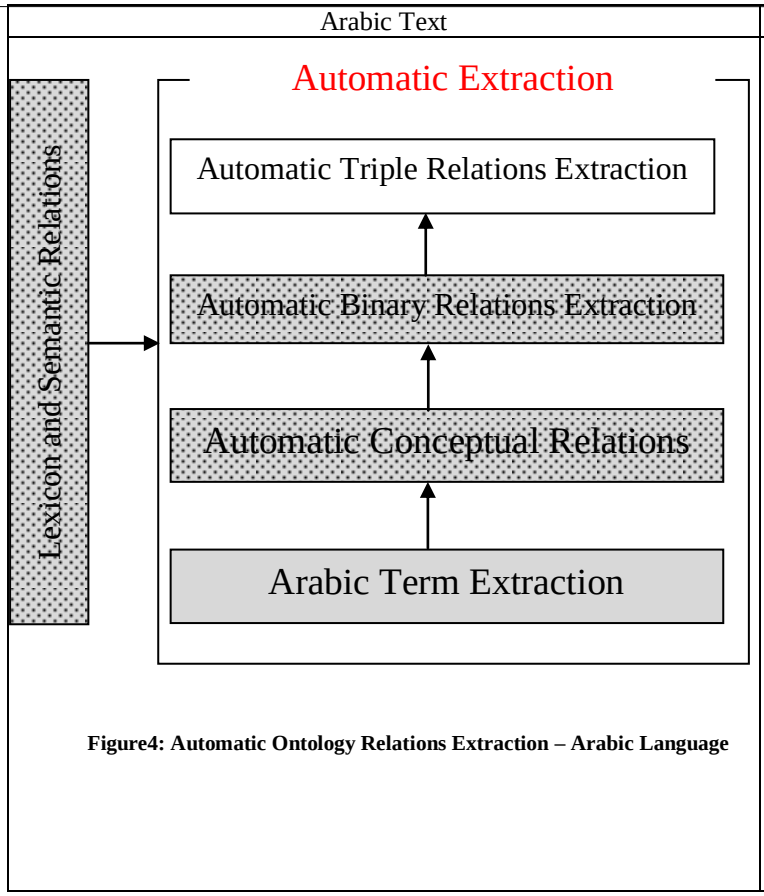

Figure4: Automatic Ontology Relations Extraction – Arabic Language

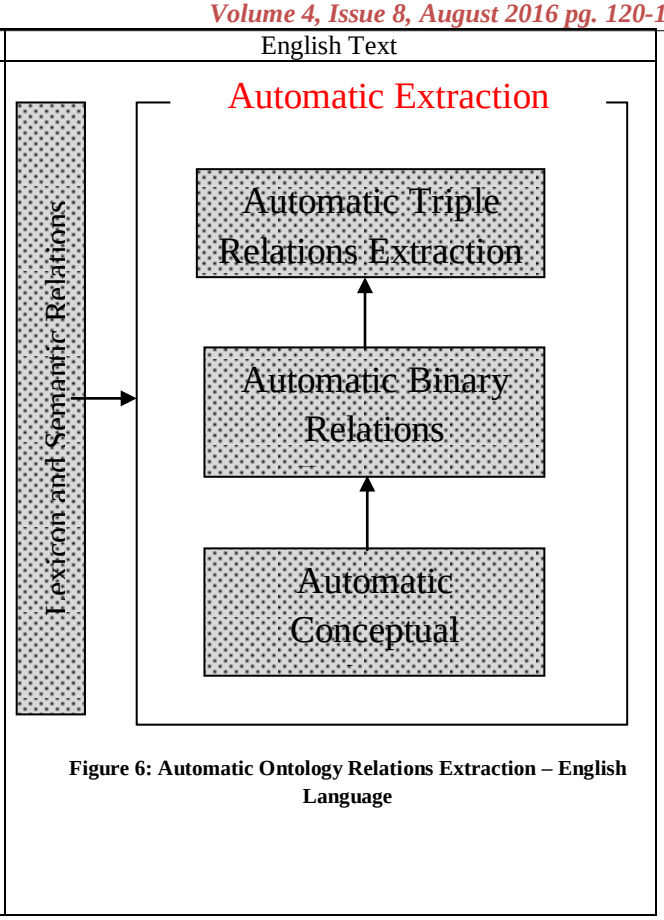

Figure 6: Automatic Ontology Relations Extraction – English Language

References

1. Belal, Mohamed "An Arabic Ontology Engine for Domain Concepts Construction and Semantic Retrieval",Paper For ITIDA.
2. Hassan, Mohsen. "A Model for Ontology Base Representation thesis." (2012).
3. Lazhar, Farek and TliliGuiassaYamina. "Identification of Opinions in Arabic Texts Using Ontologies." Information Technology & Software Engineering. (2012).
4. Arthurbrewster, Christopher. "Mind the Gap: Bridging from Text to Ontological Knowledge." (2008).
5. Yasser, Saber "An Ontology Base for Holly Quran", Graduation Project.
6. Omar Salah El-Radie, " SPARQL2AL: Translating SPARQL Queries to Arabic Language", Islamic University-Gaza, 2015.
7. Alaa W. AbuTaha, "An Ontology-Based Arabic Question Answering System", Islamic University-Gaza, 2015.
8. Saeed Al-Bukhitan, Tarek Helmy, Mohammed Al-Mulhem, " Semantic Annotation Tool for Annotating Arabic Web Documents", Saudi Arabia, 2014.
9. Wiem Lahbib, Ibrahim Bounhas, Bilel Elayeb, Fabrice Evrard, Yahya Slimani, "A Hybrid Approach for Arabic Semantic Relation Extraction", 2013.
10. Maha Al-Yahya, Sawsan Al-Malak, Luluh Aldhubayi, " Ontological Lexicon Enrichment: The Badea System for Semi-Automated Extraction of Antonymy Relations from Arabic Language Corpora", Saudi Arabia, 2016.
11. Ahmed Cherif Mazari , Hassina Aliane, and Zaia Alimazighi, Electrical Engineering and Computer science Department, University of Médéa, "Automatic construction of ontology from Arabic texts", 2012.
12. Maha Al-Yahya, Mona Al-Shaman, Nehal Al-Otaiby, Wafa Al-Sultan, Asma Al-Zahrani, Mesheal Al-Dalbahie, Information Technology Department, College of Computer & Information Sciences, King Saud University, Riyadh, Saudi Arabia, "Ontology-Based Semantic Annotation of Arabic Language Text", I. J. Modern Education and Computer, 2015.
13. Hossam Ishkewy, Hany Harb, and Hassan Farahat, "Azhary: An Arabic Lexical Ontology", Azhar University, Faculty of Engineering, Computers and Systems Engineering Department, International Journal of Web & Semantic Technology (IJWesT) Vol.5, No.4, October 2014 .
14. Mohammed G.H. Al Zamil, and Qasem Al-Radaideh, Department of Computer Information Systems, Yarmouk University, Irbed, Jordan, " Automatic extraction of ontological relations from Arabic text", Journal of King Saud University –Computer and Information Sciences, 2014.
15. Ali Mashaan Abed, Sabrina Tiun, Mohammed Albared, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Malaysia, " Arabic Term Extraction using combined approach on Islamic Document", Journal of Theoretical and Applied Information Technology, 31st December 2013. Vol. 58 No.3.
16. G. Suresh kumar, and G. Zayaraz , "Concept relation extraction using Naïve Bayes classifier for ontology-based question answering systems", Journal of King Saud University –Computer and Information Sciences, India, 13 March 2014.

17. Raghu Anantharangachar, Srinivasan Ramani, S Rajagopalan, "Ontology Guided Information Extraction from Unstructured Text", International Institute of Information Technology, Electronics City, Hosur Road, Bangalore 560100, India, International Journal of Web & Semantic Technology (IJWesT) Vol.4, No.1, January 2013.
18. Paul Buitelaar, Michael Sintek, " Middleware for Ontology Extraction from Text", Germany, 2005.
19. Paul Buitelaar, Daniel Olejnik, Michael Sintek, "A Protégé Plug-In for Ontology Extraction from Text Based on Linguistic Analysis", Germany, 2005.
20. Paul Buitelaar, Daniel Olejnik, Michael Sintek, "OntoLT: A Protégé Plug-In for Ontology Extraction from Text", DFKI GmbH, Saarbruecken/Kaiserslautern, Germany, 2004
21. k
22. Andreia Dal Ponte Novelli and José Maria Parente de Oliveira, "Simple Method for Ontology Automatic Extraction from Documents", (IJACSA) International Journal of Advanced Computer Science and Applications, 2012.
23. Hoifung Poon and Pedro Domingos," Unsupervised Ontology Induction from Text", Department of Computer Science & Engineering, University of Washington, 2010.
24. Peter Exner and Pierre Nugues, " Entity Extraction: From Unstructured Text to DBpedia RDF Triples", Department of Computer science, Lund University, 2013.
25. Maedche, Navligli, Blake and Pratt, " Automatic Extraction of a Medical Ontology From Wikipedia", 2008.
26. Alani, Harith; Kim, Sanghee; Millard, David E.; Weal, Mark J.; Hall, Wendy; Lewis, Paul H. and Shadbolt, Nigel R. (2003). "Automatic ontology-based knowledge extraction from web documents". IEEE Intelligent Systems, 2000.
27. Harith Alani, Sanghee Kim, David E. Millard, Mark J. Weal, Wendy Hall, Paul H. Lewis, Nigel Shadbolt, "Web based Knowledge Extraction and Consolidation for Automatic Ontology Instantiation", I.A.M. Group, ECS Dept. University of Southampton, Southampton, UK, 2004.
28. Stefan Bischof, 0327033, " Automatic Ontology-Based Knowledge Extraction from Web Documents", SS2005.
29. Mithun Balakrishna, Munirathnam Srikanth, " Automatic Ontology Creation from Text for National Intelligence Priorities Framework (NIPF)", Lymba Corporation, Richardson, TX, 75080, USA, 2009.
30. Alexander Maedche and Raphael Volz_, "The Ontology Extraction & Maintenance Framework Text-To-Onto", Institute AIFB, University of Karlsruhe, 76128 Karlsruhe, 2002.
31. Antonio Moreno and Chantal Pérez, "From Text to Ontology: Extraction and Representation of Conceptual Information", Conférence TIA-2001,Nansy, 3 et 4 mai 2001, Málaga, Spain, 2001.
32. Jionghua Ji, "Semi-Automatic Ontology-based knowledge extraction and verification from unstructured document ", Master of Science, Florida, 2000.
33. http://language.worldofcomputing.net/
34. John Philip McCrae, Doctor of Philiosophy, "Automatic Extraction of Logically Consistent Ontologies from text", 2009.

AUTHOR(S) PROFILE

Yasser Saber, MSc student /Computer Science Department of Faculty of computers and information, Helwan University Cairo, Egypt.

Graduated 2009 from Computer Science Department of Faculty of computers and information, Helwan University Cairo, Egypt.