



The 7th International Conference on Ambient Systems, Networks and Technologies
(ANT 2016)

Semantic Annotation of Arabic Web Resources using Semantic Web Services

Saeed Albukhitan, Ahmed Alnazer, Tarek Helmy*

*Information and Computer Science Department,
College of Computer Science & Engineering,
King Fahd University of Petroleum & Minerals,
Dhahran 31216, Mail Box 413, Saudi Arabia,
[albokhitan,alnazera,helmy]@kfupm.edu.sa*

Abstract

The vision of semantic Web is to have a Web of things instead of Web of documents in a form that can be processed by machines. This vision could be achieved on the existing Web using semantic annotation based on common and public ontologies. Due to exponential growth and huge size of the Web sources, there is a need to have fast and automatic semantic annotation services of Web documents. Since Arabic language received less attention in semantic Web research as compared to Latin languages especially in the field of semantic annotation. This motivates us in this paper to present semantic Web services that support the semantic annotation of Arabic language documents. The services accept documents and ontologies and produce annotations of these documents using different output formats. The proposed services could be used for building semantic Web applications and semantic search engines for Arabic Language. To evaluate the performance of these services, a set of ontologies were used with pre-annotated documents related to those ontologies. The initial results show a promising performance which will support the research in the semantic Web with respect to Arabic language.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Peer-review under responsibility of the Conference Program Chairs

Keywords: Semantic Web Services; Semantic Annotation; Arabic Language; Ontology, OWL-S

* Corresponding author. Tel.: +966-13-860-1967; fax: +966-13-860-2174.

E-mail address: helmy@kfupm.edu.sa

1. Introduction

Web services are application components that can be accessed via the Web programming interface. They provide functionality of a Web information system using standard technologies. However, their common descriptive languages such as WSDL¹, lack semantic richness in order for machines to process them automatically. They require human intervention to interpret their meanings for discovery, composition, and invoking. Manual intervention is an error-prone and time-consuming task. W3C supports the use of software agents for automating the above tasks. An agent is defined as application software acting on behalf of a person, a system or an organization². Accordingly, an agent could discover, compose and invoke Web services. In order for agent to perform those tasks, it requires a reference specification that includes domain informational knowledge, and operational knowledge of how to perform domain tasks. Ontology is an effective way to provide such specification.

Semantic Web service uses ontologies as their data models which bring the benefits of semantics to the executable part of the Web³. It resolves ambiguous description of service functionality and external interface. It reduces human intervention while integrating services in Service-Oriented Architecture (SOA). With semantic Web services, many tasks in the process of using Web services can be automated. SOA dynamism is improved when new services available for use as they appear. There is no need for service consumers and producers to know of each other's existence. Also, stability of the service could be improved because service interfaces are not tightly integrated so even less impact from changes. Services can be easily replaced if they are no longer available.

The amount of research and work done for annotating Arabic content in Web is very limited and non-scalable. One of the biggest challenges facing Arabic research is the availability and accessibility of Arabic resources, such as ontologies, corpora, named list, dictionaries, and NLP tools. This challenge makes collection, analyzes and investigation of such resources laborious especially if the semantic annotation techniques depend on such resources.

In this paper, we present semantic Web services that annotate the Arabic Web resources and produce annotation in different formats. The services support parsing of ontologies stored in different formats including RDF, OWL and N-TRIPLE. The services include document and ontology handling, in addition to entity and relation extraction.

The rest of this paper is organized as follows. Section 2 reviews the existing related work on semantic annotation and Web services for Web sources. Section 3 presents the proposed services. Section 4 presents the performance of the services prototype with the discussion of experiments and results. Finally, we summarize the paper and highlight the future work directions in Section 5.

2. Related Work

A survey of some semantic Web technologies supporting Arabic is presented in Beseiso et al.⁴. Four mostly used semantic Web tools were investigated, namely Protégé, Jena, Sesame, and KOAN. Their investigation focuses in the tool's functionality, type of standards supported and support level of Arabic language. Their conclusion is that those tools do not support Arabic language completely as compared to Latin languages. The most supporting tool for Arabic language was Jena with some limited support for query processing. Arabic language does not get the same support as compared to Latin languages. The common challenges of Arabic language with respect of NLP tasks were highlighted in Abdel Rahman et al. [5]. Arabic language does not have features such as case-sensitivity which is an importance feature used by Latin languages to detect proper names. Arabic words could have more than one affix and can be expressed as combination of affix such as prefixes, lemma and suffixes which make it more difficult for stemming. Arabic words also have diverse types of ambiguities associated to typographic forms and spelling.

The first Arabic semantic annotation tool was presented by Bin Saleh and Alkhalifa⁶. The presented tool was named AraTation to annotate Arabic news in the Web. The tool is capable of extracting named entities using Arabic location ontology built for this purpose. The tool reported an achievement in average precision of 67% and recall of 82% on a set of ten locations over 25 Web documents. We were not able to test the tool since the tool is not publicly available. Another work on Arabic language annotation was presented by Zaidi et al.⁷ using GATE⁸ NLP toolkit. Their system used crescent Quranic Corpus as an input. The system is capable to extract named entities through predefined patterns that use tokenized and morphology analyzed corpus with Part Of Speech (POS) features. Another semantic annotation tool for Arabic sources is presented by Motasem et al.⁹. It was tested on news article

corpus of Arabic language collected from two sources. We were not able to test the performance of these tools on our dataset for comparative analysis. We can only compare our tool with those tools based on the reported features and published performance. El-ghobashy et al.¹⁰ proposed two frameworks for semantic annotation of Web documents. In the first framework, an annotation server is being requested by proxy Web server in behave of the client. In the second framework, the annotation is done by plugging-in the tool inside a Web browser. The core of their annotation system consists of three main modules namely text preprocessing module, semantic annotation module and annotation management module. Al-Yahya et al.¹¹ presented SemTree ontology which is ontology for lexical semantic relation annotations for Arabic text. They developed a prototype system to evaluate the usefulness of SemTree ontology. Our work fills the gaps in related works and focuses on the task of facilitating annotation of Arabic text as Web services using arbitrary ontologies with different documents and ontology formats. The Web services of semantic annotation of Arabic Web documents could provide improved and automated search capabilities.

3. Annotation of Arabic Web Documents Using Semantic Web Services

In this section, we will present the architecture of the proposed semantic Web services for annotating Arabic Web documents. The services include the ontology pre-processing, document NLP analysis, entity extraction and relation extraction. Each atomic process produces an output then sent to the user or passed to the next process till the final process in the annotation pipeline. Services could produce different output format based on the intended use.

There are different formalisms available for the description of semantic Web services. A detail companions between common formalisms can be found in Kamaruddin et al.¹². In this work, we used OWL-S formulize due its simplicity and acceptability of its tools. OWL-S borrows ontological operators from OWL by providing a service upper ontology to describe Web services and service processes in a standard manner. It defines service description in three ontologies: service profile, service process and service grounding.

The *service profile ontology* describes what the service does, and is intended mainly for the purpose of service discovery. The *service process ontology* describes the composition of services that is the controlled enactment of constituent processes with respective communication pattern. The *service grounding ontology* provides a binding between the logic-based and XML-based service definitions for the purpose of facilitating service execution. Figure 1 shows the upper ontology of the Arabic semantic Web service that integrates the three sub-ontologies.

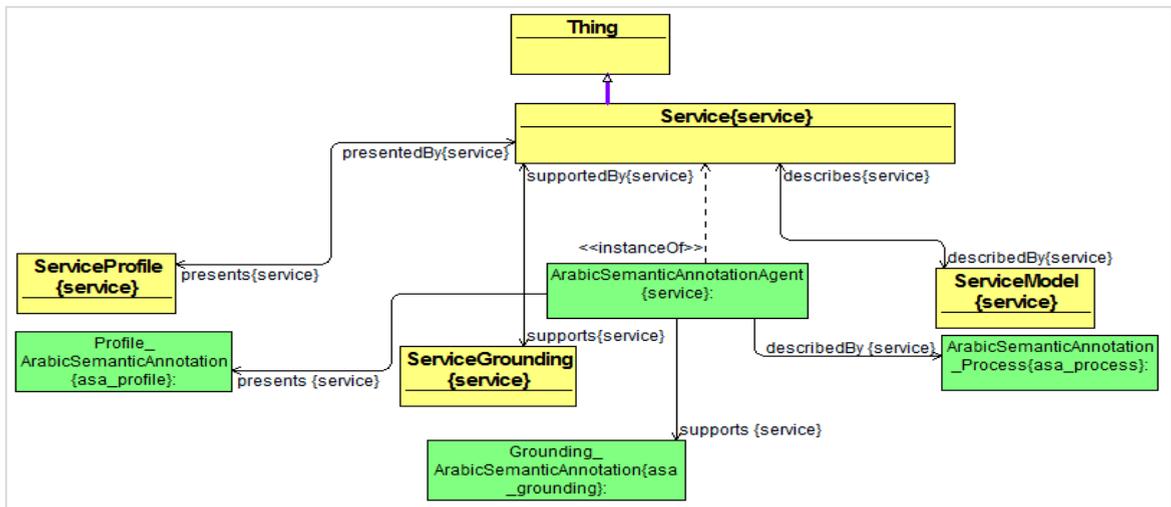


Figure 1: Upper Ontology of the Semantic Web Service

3.1. General Overview of Annotation Services

There are four services provided in the semantic Web annotation pipeline. The first service provides ontology handling and preprocessing function to make the ontology useable for annotating Arabic documents. The second service provides the handling of documents provided by users or applications. Document handling includes the format handling and extracting useful chunk of text in the document. It also performs basic and advanced NLP analysis of input documents. The third service is responsible for extracting named entities found in the input documents with respect the used ontologies. The forth service is responsible for discovering relationship between the recognized named entities in the same document. A composite process that makes use of the atomic processes could be seen in Figure 2. Next, an overview of each service will be given with the type of an input and output parameters. Sample of output will be shown also.

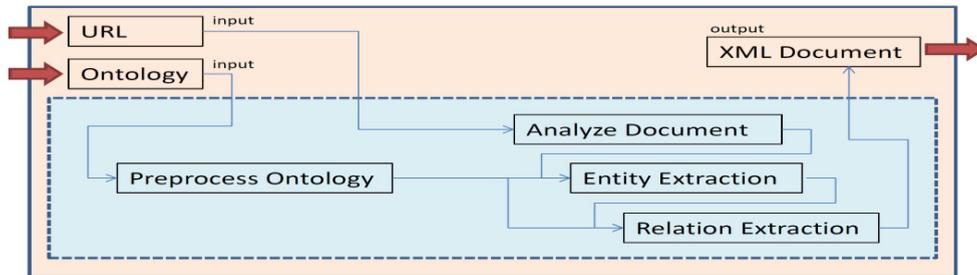


Figure 2: Process Model for the Composite Process of Annotating a URL

3.2. Ontology Handling Semantic Web Service (OHSWS)

The Ontology Handling Semantic Web Service (OHSWS) prepares the provided ontology for the annotation process. In order to annotate a text against a given ontology, constructs in the ontology need to be extracted and matched with the text chunks. The matching process differs for different constructs, Ontology constructs need to be augmented with additional lexical such as synonym of the label and name of ontology concepts. External resources are used for this purpose such as WordNet and dictionaries. Description of OHSWS service is shown in Table 1 and Figure 3.

OHSWS service provides four atomic processes. The *GetAugmentedOntology* process enriches a given ontology with additional information gather from external resources such as WordNet and Wikipedia. It also provides the one2one translation of concepts into Arabic language in case of the provided ontology constructs are not in Arabic. The *GetEntityDictionary* process builds a dictionary of the ontology concepts with synonym with different data structure for quick processing of document. The *GetRelationDictionary* process builds a dictionary of relationship between ontology concepts with additional constraints. The last *GetCompleteDictionary* process is a composite process that combines all the other processes and builds a comprehensive data structure for the input ontology.

Table 1: Description of the Ontology Handling Semantic Web Service (OHSWS)

Process	Process Details	
GetAugmentedOntology	Input Parameters	Ontology in OWL format (required parameter)
		Use external resources: Yes, No (default)
	Output Parameters	Ontology in OWL format
GetEntityDictionary	Input Parameters	Ontology in OWL format (required parameter)
		Use external resources: Yes, No (default)
	Output Parameters	Dictionary Xml
GetRelationDictionary	Input Parameters	Ontology in OWL format (required parameter)
		Use external resources: Yes, No (default)
	Output Parameters	Dictionary Xml

GetCompleteDictionary	Input Parameters	Ontology in OWL format (required parameter)
		Use external resources: Yes, No (default)
	Output Parameters	Dictionary in XML format

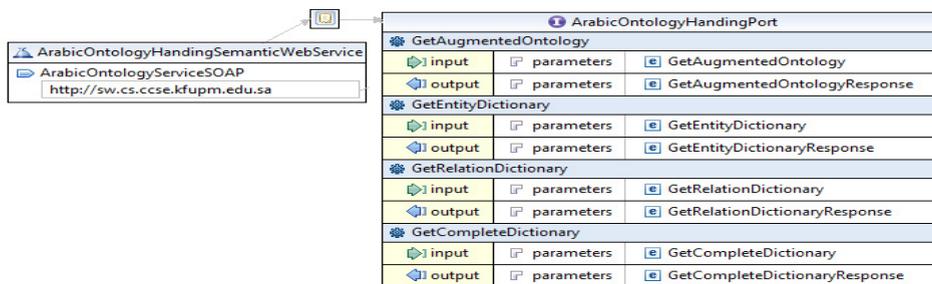


Figure 3: WSDL definition of Ontology Handling Semantic Web Service (OHSWS)

3.3. Document Analysis Semantic Web Service (DASWS)

Document Analysis Semantic Web Service (DASWS) provides format handling and extraction of useful chunk of text in the input document while discarding tags and non-useful parts. It also performs NLP analysis of the document. In addition to the tokenization and normalization, it uses NLP analyses steps of Arabic text in order to collect the features that are needed for entity and relationship extraction. *Part-of-speech* (POS) task makes use of an open source POS tagger such as Stanford inside GATE tool. *Parsing* task uses Arabic Parser such as Stanford Parser with Arabic configuration to generate a parse tree for Arabic sentences. *Morphological Analysis* task makes use of publicly available and open source java version of Buckwalter Arabic Morphological Analyzer to get the root of Arabic works. *Stemming* is required since named entities in Arabic language have attached prepositions and conjunctions often. It is quite useful in analyzing named entities by removing those additives making job easy for search-based applications to function. We have used an open stemmer provided by Khoja¹³.

DASWS service provides two atomic processes. The *GetNLPProcessedText* process performs a complete NLP analysis pipeline for a given document. The *GetDocumentMetaData* process provides the meta-data and some aggregation values for requested document. Description of OHSWS service is shown in Table 2 and Figure 4.

Table 2: Description of the Document Analysis Semantic Web Service (DASWS)

Process	Process Details		
GetNLPProcessedText	Input Parameters	Document URL (required parameter)	
		POS tagger	Stanford [25], APT [26], QCRI [27]
		Parser	Stanford, Bikel, Berkeley Parser
		Stemming	No (default), Yes
	Output Parameters	Xml	
GetDocumentMetaData	Input Parameters	Document URL (required parameter)	
	Output Parameters	Xml	

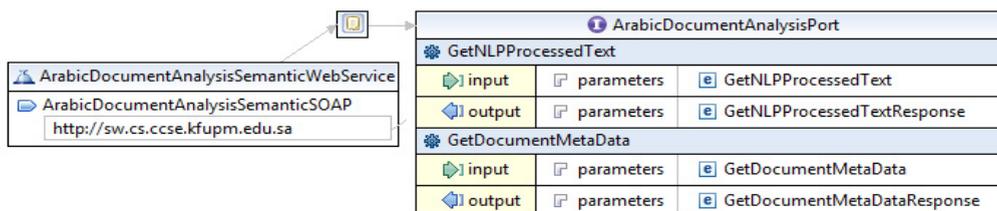


Figure 4: WSDL Definition of Document Analysis Semantic Web Service (DASWS)

3.4. Entity Extraction Semantic Web Service (EESWS)

Entity Extraction Semantic Web Service (EESWS) is used for the identification of named entities such as people, organizations, events, companies, cities, geographic features, and other typed entities within the given Web content. It makes use of a combination of rule-based and machine learning algorithms after Natural Language Processing (NLP) technology. NLP is applied to analyze textual information in order to extract the semantic richness embedded within Web content. For named entities recognition, fuzzy string matching is used to address the problem of spellings variation in Arabic instance names. For fuzzy string matching, EESWS uses ensemble of string matching algorithms with difference weights to calculate the matching between two strings.

EESWS service provides two atomic processes. The *GetNamedEntitiesURL* process extracts grouped, relevancy-ranked list of named entities (people, companies, organizations, etc.) from a given URL. EESWS service will download the requested URL, extract text from the document structure (ignoring navigation links, advertisements, and other undesirable content), and performs entity extraction operations. The *GetNamedEntitiesDocument* performs the same function of *GetNamedEntitiesURL* process with the exception that it will receive the document part of input parameters. Description of EESWS service is shown in Table 3 and Figure 5.

Table 3: Description of the Entity Extraction Semantic Web Service (EESWS)

Process	Details		
GetNamedEntitiesURL	Input Parameters	Document URL	(required parameter)
		OntologyMode	
		OutputFormat	xml (default), json, or rdf
	Output Parameters	xml , json, or rdf document	
GetNamedEntitiesDocument	Input Parameters	Document file(required parameter)	
		OntologyMode	UseDBpedia (default), UsePredefinedSet, UserProvided
		OutputFormat: xml (default), json, or rdf	
	Output Parameters	xml, json, or rdf document	

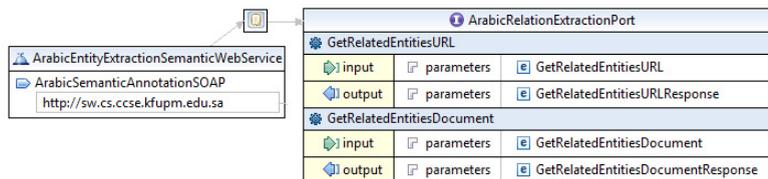


Figure 5: WSDL Definition of Entity Extraction Semantic Web Service (EESWS)

3.5. Relation Extraction Semantic Web Service (RESWS)

Relation Extraction Semantic Web Service (RESWS) is used for the identification of relationship between entities using different relation templates such as Subject-Action-Object relations within Web content. It makes use of a combination of rule-based and machine learning algorithms to analyze textual information in order to extract the semantic richness embedded within Web content. The rule-based extraction is based on grammar rules using GATE JAPE which provides Finite State Transduction (FST) based on regular expressions over annotations. JAPE uses a syntax that contains of a set of phases. Each phase contains of a set of pattern with action rules. The phases run in a sequential manner and establish a cascade of FST over annotations. Rules in the Left-Hand-Side (LHS) contain of an annotation pattern description while the rules in Right-Hand-Side (RHS) consist of list of statements that manipulate annotations using either Jape syntax or Java code. Pattern elements in matched annotations on the LHS of a rule may be attached with labels. The RHS statements could refer to those labels to manipulate annotations.

RESWS service provides two atomic processes. The *GetRelatedEntitiesURL* process detects the relationship between recognized entities in the provided document. It will verify the detected relationships with ontology

constructed for localized validation. The *GetRelatedEntitiesDocument* performs the same function of *GetRelatedEntitiesURL* process with the exception that it will receive the document part of input parameters. Description of EESWS service is shown in Table 4 and Figure 6.

Table 4: Description of the Relation Extraction Semantic Web Service (RESWS)

Process	Details		
GetRelatedEntitiesURL	Input Parameters	Document URL	(required parameter)
		OntologyMode	UseDBpedia (default) , UsePredefinedSet, UserProvided
		OutputFormat	xml (default), json, or rdf
	Output Parameters	xml , json, or rdf document	
GetRelatedEntitiesDocument	Input Parameters	Document file(required parameter)	
		OntologyMode	UseDBpedia (default) , UsePredefinedSet, UserProvided
		OutputFormat: xml (default), json, or rdf	
	Output Parameters	xml, json, or rdf document	

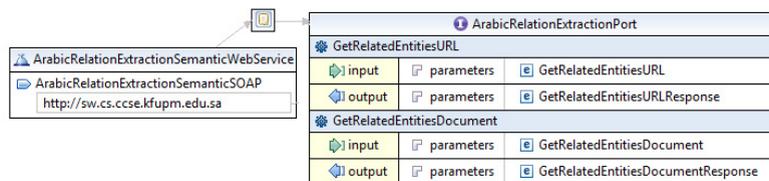


Figure 6: WSDL Definition of Relation Extraction Semantic Web Service (RESWS)

4. Prototype Evaluation and Discussion

In order to evaluate the effectiveness of the provided services, we developed a semantic annotation tool that makes use of the provided services. The tool has been developed using Java. We have used a set of documents related to three domains: Food, Nutrition and Health. We have used ontologies for those three domains containing 8012 food concepts, 146 nutrition concepts and 4516 health concepts. The documents set contain 150 HTML files and annotated by hand according to domain ontologies. Then, we compared the annotated information by the developed tool to the reference set to evaluate the performance of the extraction process using the Precision, Recall and F-Measure metrics:

- **Precision** is the percentage of correctly recognized information from the total number recognized information.
- **Recall** is the percentage of information in the reference set that were recognized.
- **F-measure** is a harmonic mean of precision and recall, given by: $F\text{-measure} = \frac{2 * recall * precision}{recall + precision}$

The summary of the results obtained from the tool is shown in Table 5. We have incorporated the Precision, Recall, and F-Measure values for each named entity category. Nutrition concepts annotation had the lowest performance in both Precision and Recall. This could be due to the complexity of this type of named entities representation in Arabic language. Base on this evaluation, the results demonstrate a promising Precision and Recall.

Table 5: NE Recognition Performance

Ontology	Entity Concepts	Total NEs	Extracted	Correct	Wrong	Precision	Recall	F-Measure
Food	Food Items	98	82	71	11	86.6%	72.4%	78.9%
Nutrition	Nutrients	187	147	123	24	83.7%	65.8%	73.7%
Health	Disease	200	163	137	26	84.0%	68.5%	75.5%

Table 6 summarizes the recognition accuracy, in terms of Precision, Recall, and F-measure, achieved by the developed tool for facts in term of the relationships between food/nutrition and disease concepts.

Table 6: Food/Nutrition and Disease Relation Extraction Performance

Relation	Total Relations	Extracted	Correct	Wrong	Precision	Recall	F-Measure
Food- Disease	43	37	31	6	83.80%	72.10%	77.50%
Nutrition - Disease	52	50	39	11	78.00%	75.00%	76.50%

We have shown in this paper a limited evaluation of the functional performance of the tool with respect to entity and relation extraction. Also, we have done many performance evaluations for each service with respect to expected outcome that was done by hand. After a lot of enhancement to each service, we reach to an acceptable performance level. There are still area of performance improvement and scope in each service that will be shown in future publication.

5. Conclusion and Future Work

This paper presents the development of a semantic Web service for semantic annotation of Arabic Web resources. The novelty of this work resides in leveraging semantic Web technologies to serve the Arabic language, and produces semantically annotated Web documents for the targeted domains in an automatic manner. The achieved performance is promising. There are still many potential extensions that can enhance the services performance and output. We are planning to improve the performance by integrating the services with additional lexical resources and tools. We would like to include additional public ontologies such as Freebase and Yago after having Arabic support in them.

Acknowledgements

The authors would like to thank King Fahd University of Petroleum and Minerals for supporting this work through the project no. IN141038. In addition, the authors would like to thank both conference Chairs and the anonymous reviewer's for their valuable comments that enhance the paper presentation.

References

- Christensen, E., Curbera, F., Meredith, G., Weerawarana, S., "Web Services Description Language (WSDL), Version 1.1", W3C Note, 15 March 2001. <http://www.w3.org/TR/wsdl>, accessed 2015-12-22.
- Hugo Haas, and Allen Brown. "Web Services Glossary", W3C Working Group Note 11 February 2004, <http://www.w3.org/TR/ws-gloss/>, accessed 2016-01-02.
- Carlos Pedrinaci, John Domingue, Amit P. Sheth, "Semantic Web Services", Handbook of Semantic Web Technologies, pp. 978-1037, Springer, 2011
- M. Beseiso, A. R. Ahmad, and R. Ismail, "A Survey of Arabic language Support in Semantic web", Int. J. Comput. Appl., vol. 9, no. 1, pp. 35–40, Nov. 2010.
- S. AbdelRahman, M. Elarnaoty, and M. Magdy, "Integrated Machine Learning Techniques for Arabic Named Entity Recognition," Int. J. Comput. Sci. Sci., vol. 7, no. 4, pp. 27–36, 2010.
- L. M. Bin Saleh and H. S. Al-khalifa, "AraTation: An Arabic Semantic Annotation Tool," Proc. Int. Conf. Inf. Integr. Web-based Appl. Serv., pp. 447–451, 2009.
- S. Zaidi, M.-T. Laskri, and A. Abdelali, "Arabic Collocations extraction using Gate," Int. Conf. Mach. Web Intell., pp. 473–475, 2010.
- H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters, Text Processing with GATE (Version 6). The University of Sheeld, Department of Computer Science, 2010.
- Alrahabi Motasem, Ibrahim Amr Helmy, Desclés Jean-Pierre, "Semantic Annotation of Reported Information in Arabic", FLAIRS 2006, Melbourne, Florida.
- El-ghobashy, Ahmed N., Gamal M. Attiya, and Hamdy M. Kelash. "A Proposed Framework for Arabic Semantic Annotation Tool." Int. J. Com. Dig. Sys 3.1 (2014): 47-53.
- Maha Al-Yahya, Mona Al-Shaman, Nehal Al-Otaiby, Wafa Al-Sultan, Asma Al-Zahrani, Mesheal Al-Dalbahie, "Ontology-Based Semantic Annotation of Arabic Language Text", IJMECS . July 2015, Vol. 7 Issue 7, p53-59. 7p.
- Lina Azleny Kamaruddin, Jun Shen, and Ghassan Beydoun, "Evaluating usage of WSMO and OWL-S in semantic web services" . In Proceedings of the Eighth Asia-Pacific Conference on Conceptual Modelling - Volume 130 (APCCM '12), 2012.
- <http://zeus.cs.pacificu.edu/shereen/research.htm>