MATTER: International Journal of Science and Technology
ISSN 2454-5880

Hendi et al., 2018

Volume 4 Issue 3, pp. 87-96

Date of Publication: 19th November, 2018

DOI-https://dx.doi.org/10.20319/mijst.2018.43.8796

This paper can be cited as: Hendi, H. G., Al-Feel, H. & Hassanein, E. E. (2018). Annobıc Annotatıon of Arabıc RSS Feed. MATTER: International Journal of Science and Technology, 4(3), 87-96.

ANNOBIC ANNOTATION OF ARABIC RSS FEED

Hanaa Ghareib Hendi
Faculty of Computers & Information, Fayoum University, Egypt
hgm01@Fayoum.edu.eg

Haytham Al-Feel
Faculty of Computers & Information, Fayoum University, Egypt
htf00@Fayoum.edu.eg

Ehab E. Hassanein
Faculty of Computers & Information Cairo University, Cairo, Egypt
e.ezat@fci-cu.edu.eg

Abstract

Annotation is adding metadata to pages to become more meaningful and readable for machines. However, many Semantic annotation tools developed which proved their success in multiple languages, but Arabic is none of them. We present AnnoBic which is an Arabic semantic annotation tool for RSS feeds.

Keywords

Semantic Annotation, Arabic Language, RSS, Ontology

1. Introduction

The current web has many problems and limitations. Till now most of web pages can't being readable by machines, they are available for humans which makes the web flat and boring (T.Berners-Lee,M. Fischetti, and M.L. Dertouzos, 2000).The semantic web is not a new version of the web but an extension of the current one aims to improve the current existing web with an extra machine interpretable layer with metadata. This metadata allows the machines to understand what a web page is about, which makes data easier to be shared and exchanged and knowledge to be accessible and easy to discover.

Adding metadata to resources we called it "annotation". Semantic annotation is the process of tagging web content with semantics of their contents (F. van Harmelen, 2004).This process is seen as a dynamic creation of bidirectional relationships between ontologies and web documents (Bontcheva, Kalina, and Hamish Cunningham, 2011).

Ontology as we discussed earlier one of the main players for the semantic web but also for annotations. It is commonly defined as an explicit formal specification of a shared conceptualization of a domain of interest (Thomas R.Gruber, 1995). Our paper presents a tool that provides Semantic annotation to Arabic News web content by reading RSS (Really Simple Syndication) feeds and match instances with ontology classes. RSS is an XML application that shares updated information where interested users can collect and subscribe (K.Holvoet, 2006).

The remainder of this paper is organized as follows: Section 2, we present the importance and challenges of Arabic language. Afterward, we preview a background of Annotation Tools that support Arabic language in Section 3. In section 4, we focus on our RSS Annotation tool (i.e. AnnoBic) and discuss architecture, algorithm. In Section 5, we evaluate AnnoBic Annotation tool and discuss the results. Finally, in section 6 presents our conclusions and further research.

2. Semantic Web and Arabic Language

The importance of the Arabic language comes from more than two hundred sixty million speak it. There are many difficulties face developers to implement semantic tools in Arabic language. Part of these difficulties return to the language itself and another part return to the availability of ontologies in Arabic let's discuss these difficulties briefly:-

- Arabic language loss capital letter which is an important feature to identify proper Entity since Named Entity (NE) usually start with capital letters.

- Arabic word can be expressed as a combination of prefix(s), lemma, and suffix(s) that make it hard for stemming (A. Saeed, 2008).
- Arabic language is ambiguity; the same text has a different significance. Arabic has a typographic variance. For example "امريكا-أمريكا"/"America".
- There are rare in Ontologies, Corpus, and NLP tools that written to Arabic uses. This imperfection makes gathering and analyzing of different resources to be time-consuming, especially in annotations techniques depend on it. On the other hand. There are good numbers of work that have been published about Semantic Web data sources and ontologies in English (A. Al-Nazer, S. Albukhitan, and T. Helmy, 2016).

Table 1 summarizes the comparison between Arabic annotation tools and our AnnoBic tool.

Table 1: *An Arabic Semantic Annotation Tool Comparison*

Tool	Standard format	Document format	Output
Motasem et al	None	Text	Text
Zaidi et al.	OWL	Text	RDF
AraTation	OWL	HTML	RDF
AMASAT	RDF(S), OWL	HTML,TXT	RDF, RDFa
AnnoBic (our tool)	RDF,OWL	XML,TXT	RDF, RDFa

3. Arabic Semantic Annotation Tool

- **Motasem et al.** *tool* An Arabic tool for annotation of Arabic News. This tool success by carries out semantic annotation of the information search in a text that specifying the author's position, and where and when he read that. The tool based on Arabic language text corpus from two sources of press articles *(A. Motasem, I.Amr Helmy, and Desclés Jean-Pierre, 2006).*
- **Zaidi et al.** ***tool*** an Arabic annotation tool that uses Natural language Processing (NLP) GATE *(Gate, n.d.)* software toolkit. This tool depends on Quranic Corpus to be annotated through predefined patterns and kept in data store. These patterns used morphology as Part Of

Speech (POS). The evaluated tool was 0.66 for precision and recall *(Soraya Zaidi , M-T. Laskri ,and Ahmed Abdelali, 2010).*

- **AraTation** an Arabic annotation tool that semantically annotates Arabic Location Geography. This tool has the capability of extracting named entities through its location ontology. Words extracted from IE (information Extraction) mapped to the equivalent ontology instance, then the annotated documents are saved in RDF file. The obtained results in average precision of 67% and recall of 82% *(Layan M. Bin Saleh, Hend S. Al-Khalifa, 2009).*
- **AMASAT** an Automatic Arabic Semantic Annotation Tool that specialized in three domains they are: Food, Nutrition, and Health. It produces RDFa and RDF files. AMASAT maps terms in text file to the corresponding resources in the ontology. One of the advantages of AMASAT its capability to analysis the explicit and implicit relationships *(Al-Bukhitan, Saeed, Tarek Helmy, and Mohammed Al-Mulhem, 2014).*

4. System Overview

AnnoBic's architecture is presented in Figure 1.The system is divided into 4 basic components:-

- **Text Preprocessing**, which cleans the text before applying it.
- **Text Analyzing**, *which is responsible for parsing and extracting the textual* information from the web document.
- **Data Matching**, *which supplies the power to identify the similarity* between the extracted data and the ontology resources' term.
- **Semantic Annotation**, *which maps the extracted words to corresponding instances. Then the annotated document is saved in an RDF file.*

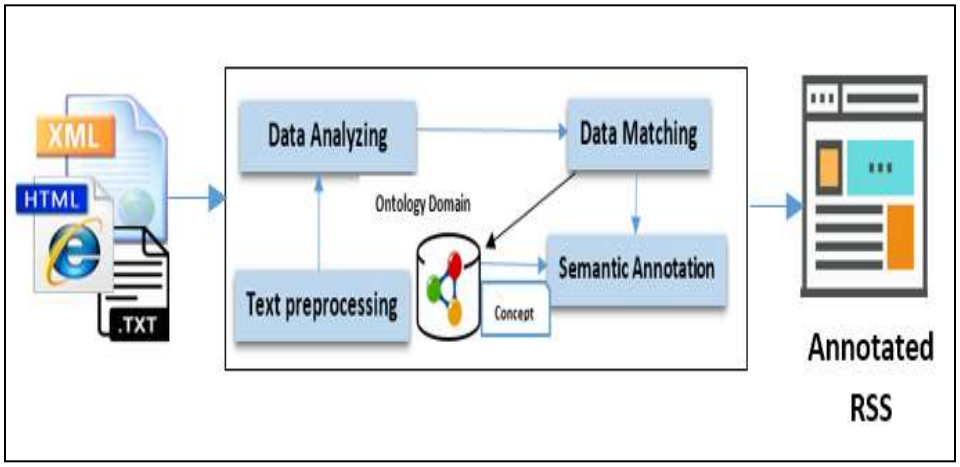


Figure 1: *AnnoBic Architecture*

In AnnoBic algorithm, Preprocess function is responsible for setting up documents before processing by any application such as:

- Replacing "أ","إ", and"آ" by "ا".
- Dismissing the effect of connector letter at the beginning of the original word like "ل","ف","ب", or "و" at "لالمانيا" wherever the original news is "المانيا".
- Neglecting any noisy from words to its original and dealing with it to achieve the best result.

We implemented AnnoBic tool as a web application using Java Programming language. We used Jena as a Java framework with programmatic environment for semantic web applications (JENA, n.d.), in addition to that we used Protégé as an ontology editor (Protégé , n.d.). According to unavailability of ontology in the Arabic language for news, we build our ontology that matched with RSS news gathered from news portal.

AnnoBic ontology presents world continents, countries, cities, rivers, oceans and their relations as shown in Figure 2.

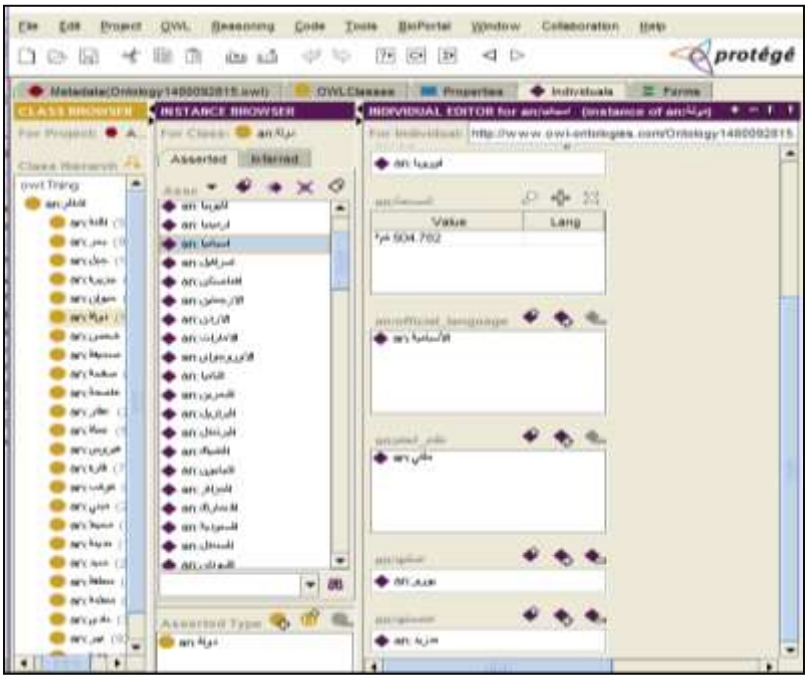


Figure 2: *AnnoBic Ontology modeling*

AnnoBic User Interface (UI) allows users to insert URLs for Arabic news RSS. After RSS URL validation, the AnnoBic tool processes the RSS news as seen in Figure 3.

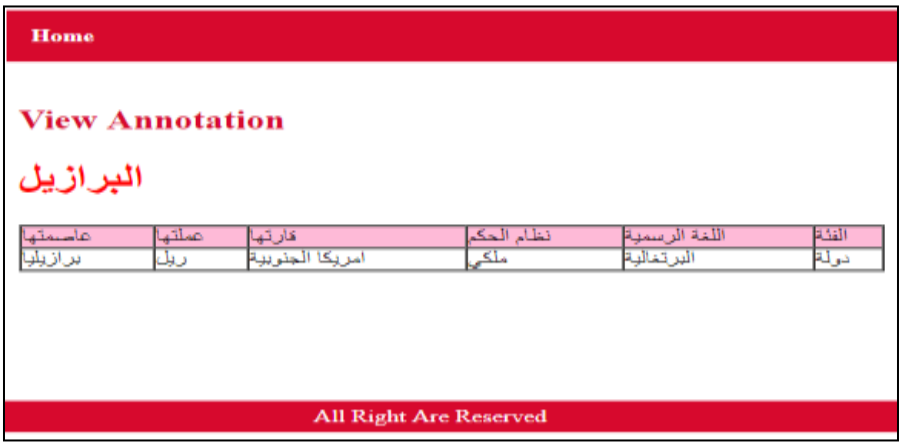
Figure 3: *AnnoBic Home Page*

AnnoBic processes RSS news feeds line by line. Every line has words with different colors based on the instance of the class represents that word as shown in Figure 4. The green color represents an annotation of a person instance, the blue color represents an annotation of an ocean instance, and violet color represents annotation is a country instance.

On the right side of the browser different categories of the named entity are presented; these named entities could be Continent, Country, Capital, and City.

Figure 4: *AnnoBic Annotation Process*

AnnoBic interface has two buttons they are "Highlight" and "view"; the first one enables users to mark important words from their point of view, while the another button enables users to view all metadata about an annotated instance. By clicking "View" button the full annotation of the selected instance extracted from the OWL knowledgebase (KB) and displayed in the form of HTML table as shown in Figure 5.



عاصمتها	عملتها	قارتها	نظام الحكم	اللغة الرسمية	الفئة
برازيليا	ريال	امريكا الجنوبية	ملكي	البرتغالية	دولة

Figure 5: *AnnoBic Full Annotation*

5. AnnoBic Evaluation

AnnoBic performance is measured by calculating the standard measures of Precision, Recall, and F-measure. We use 10 different RSS feeds to evaluate the system performance depends on bellow equations (J. Makhou, F.Kubala , R.Schwartz ,and Ralph, 1999) (Powers, David Martin, 2011). The Equations of precision and recall definition presented below.

$$\textbf{Recall} = \frac{accurate}{all} \quad (1)$$

$$\textbf{Precision} = \frac{accurate}{accurate+inaccurate} \quad (2)$$

Where "accurate" are things that annotated correctly, and "inaccurate" things annotated wrongly or not annotated before. "Accurate" and "inaccurate" annotations are generated by our RSS annotation tool, where "all" refers to all annotations supposed to be generated. The precision is a measure of correctness.

F-measure is the harmonic mean of precision and recall which refers to completeness that reflects the actual performance of the system.

$$\textbf{F-measure} = 2* \left(\frac{Percision \times Recall}{Percision+Recall}\right) \quad (3)$$

As Table 2 shows, Precision are generally 100% because the system only retrieves the ads that have a matching OWL repository with the RSS instances.

Table 2: *Precision, Recall and F-Measure of AnnoBic for 10 Different RSS Feeds*

	Precision (%)	**Recall (%)**	**F-Measure (%)**
RSS1	100	93.8	96.8
RSS2	100	92.6	96.2
RSS3	100	91.4	95.5
RSS4	100	83.6	91.1
RSS5	100	83.3	90.9
RSS6	100	82.1	90.2
RSS7	98.2	85.7	91.5
RSS8	97.6	91.9	94.2
RSS9	95.7	88	91.7
RSS10	90.9	80	85.1

For ten iterations for the AnnoBic, F-measure results range from 85% to 96%, by an average 92.3. %. While recall results ranges from 93.8 % to 80% and an average is 87.16. In

addition to that, Precision almost be 100% by an average 98.24. Figure 6 presents the relation between precision and recall and f-measure.

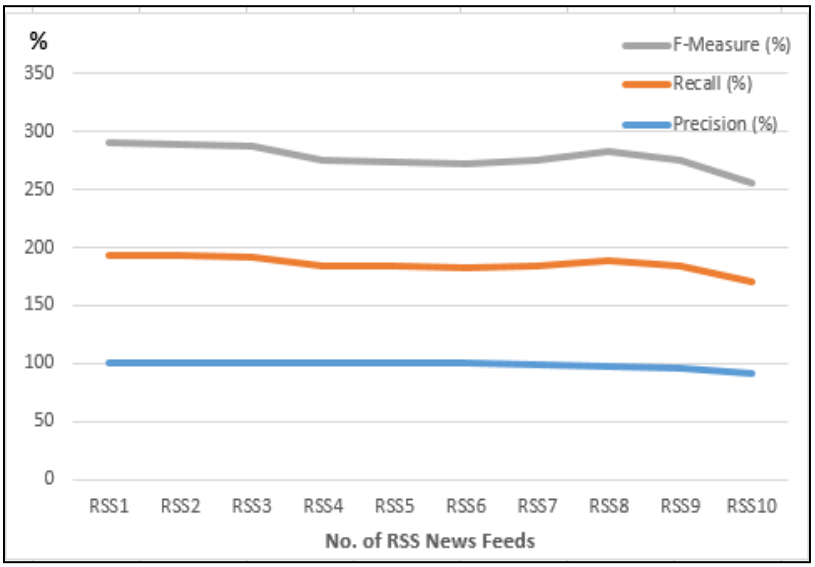


Figure 6: *AnnoBic Precision, Recall and F-measure*

6. Conclusion and Future Work

Our paper presents a short survey for semantic annotation platforms and classified their types. In addition, we presented the design, implementation, and evaluation of the AnnoBic Tool, a tool for annotating Arabic RSS news feeds semantically. Arabic Language is the top of the widest spoken language in the universe, but there is a few number of Arabic ontology in special fields.

References

A. Al-Nazer, S. Albukhitan, and T. Helmy. (2016). Cross-Domain Semantic Web Model for Understanding Multilingual Natural Language Queries: English/Arabic Health/Food Domain Use Case. *Procedia Computer Science*, 607-614.

A. Motasem, I.Amr Helmy, and Desclés Jean-Pierre. (2006). Semantic Annotation of Reported Information in Arabic. *FLAIRS.* Melbourne, Floride .

A. Saeed. (2008). *The Qur'An: An Introduction.* Routledge.

Al-Bukhitan, Saeed, Tarek Helmy, and Mohammed Al-Mulhem. (2014). Semantic annotation tool for annotating arabic web documents. *Procedia Computer Science, 32*, 429-436.

Bontcheva, Kalina, and Hamish Cunningham. (2011). Semantic annotations and retrieval: Manual, semiautomatic, and automatic generation. *Handbook of semantic web technologies* (pp. 77-116). Springer Berlin Heidelberg.

F. van Harmelen. (2004). The semantic Web: what, why, how, and when. *IEEE Distributed Systems Online, 5.*

Gate. (n.d.). Retrieved May 25, 2017, from https://gate.ac.uk.

J. Makhou, F.Kubala , R.Schwartz ,and Ralph. (1999). Performance Measures For Information Extraction. *Proceedings of DARPA broadcast news workshop.*

JENA. (n.d.). Retrieved Feb 17, 2017, from http://jena.apache.org/

K.Holvoet. (2006). What is RSS and how can libraries use it to improve patron service? . *Library Hi Tech News, 23*(8), 32-33.

Layan M. Bin Saleh, Hend S. Al-Khalifa. (2009). AraTation : An Arabic Semantic Annotation Tool. *11th International Conference on Information Integration and Web-based Applications & Services* (pp. 447–451). ACM.

Powers, David Martin. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies, 2*(1), 37-63.

Protégé . (n.d.). Retrieved May 20, 2017, from http://protege.stanford.edu/

Soraya Zaidi , M-T. Laskri ,and Ahmed Abdelali. (2010). Arabic Collocations extraction using Gate. (pp. 473–475). Algiers, Algeria: IEEE.

T.Berners-Lee,M. Fischetti, and M.L. Dertouzos. (2000). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor.* HarperInformation.

Thomas R.Gruber. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies, 43*(5-6), 907-928.