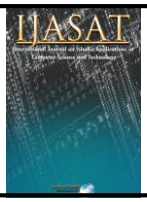

Towards Concept Extraction for Ontologies on Arabic language

Abeer Al-Arfaj [a] and AbdulMalik Al-Salman[b]

College of Computer and Information Sciences, Computer Sciences Department, King Saud University, Riyadh, Saudi Arabia
[a]abeerarfaj@yahoo.com, [b]salman@ksu.edu.sa

ABSTRACT

Ontology is one of the most popular representation model used for knowledge representation, sharing and reusing. The Arabic language has complex morphological, grammatical, and semantic aspects. Due to complexity of Arabic language, automatic Arabic terminology extraction is difficult. In addition, concept extraction from Arabic documents has been challenging research area, because, as opposed to term extraction, concept extraction are more domain related and more selective. Manual concept extraction is time-consuming process and not objective. Automatic concept extraction methods often analyze a document to determine the important domain terms, which can be a single word or multi-word term. In the literature, there are many approaches, techniques and algorithms used for term extraction. In this paper, we deal with fundamental layers involved in ontology construction from Arabic text: extracting the relevant domain terminology from a text and discovering domain concepts. Moreover, we study the problem of Arabic concept extraction from domain texts and provide a comparative review of the existing Arabic term extraction approaches highlighting the challenges posed by Arabic language characteristics. Despite the efforts to combine methods on Arabic term extraction, the field is still open for new development. The paper also proposes a future study to address this issue.

***Keywords*: Ontology Construction, Arabic Ontology, Arabic Language Processing, Concept Extraction, Arabic Term Extraction, Specific domain corpus.**

1. Introduction

Ontology is one of the most popular representation model used for knowledge representation, sharing and reusing. Ontology has been used in wide applications like knowledge management, information retrieval, information integration, bioinformatics and e-learning. Ontology construction includes several steps as follows: term extraction, synonyms extraction, concept learning, finding relations between extracted concepts and adding them in the existing ontology (Al-Arfaj and Al-Salman 2015a). Automatic extraction of concepts is one of the most important tasks of ontology learning. Term extraction is a prerequisite for all aspects of ontology learning from text. Its purpose is to extract domain relevant terms from natural language text. Terms are the linguistic realization of domain specific concepts. Term can be a single word or multi-word compound relevant for the domain in question as a term (Cimiano, 2006).

The Arabic language has complex morphological, grammatical, and semantic aspects since it is a highly derivational and inflectional language, which makes morphological analysis a very

complex task. Therefore, the NLP tools that were designed for English cannot meet the need of the Arabic language. In addition, the Arabic language lacks the capitalization feature, which makes the extraction of Arabic Named entities a complex task. The Arabic language is highly ambiguous when vowelization feature is absent (Elkateb et al., 2006; Beseiso et al., 2010; Beseiso et al., 2011; Farghaly and Shaalan 2009). Many levels of ambiguity pose a significant challenge to researchers developing NLP systems for Arabic (Attia, 2010). Researchers have found ambiguity in Arabic to be present at several levels of analysis (Farghaly and Shaalan 2009; Bounhas and Slimani 2009). Internal word structure ambiguity, that is, when a complex Arabic word could be segmented in different ways. Syntactic ambiguity, sematic ambiguity, constituent boundary ambiguity and anaphoric ambiguity. All these difficulties pose a significant challenge to researchers developing NLP systems in general and particularly on the terminologies extraction for Arabic.

To build Arabic ontology, the first step is to find the important concepts of the domain. The concept linguistically is represented by terms, so to extract the domain specific terms from texts. For English there are some studies done for concept extraction, moreover, there are some studies for unstructured Arabic documents for key phrase extraction and multiword terms extraction such as (El-Beltagy and Rafea 2008; Boulaknadel et al., 2008; Bounhas and Slimani 2009; Saif and AbAziz 2011). However, key phrase extraction is different from concept extraction. In the framework by (Al-Arfaj and Al-Salman 2014), concept extraction consists of terminology extraction and concept identification. Concept extraction from Arabic documents has been challenging research area, because, as opposed to term extraction, concept extraction are more domain related and more selective.

The main contributions of the paper are as the following:

- We provide an extensive analysis of term extraction approaches.
- We specifically summarize the Arabic terminology extraction methods, with main intent of highlighting their strengths and weaknesses on extract domain relevant terms.
- We propose a new future research direction of Arabic domain term extraction.

The rest of the paper is organized as follows. Section 2 explains concept extraction procedure and some basic definition associated with terms is presented in section 3. In section 4, we discuss existing approaches for Arabic terminology extraction followed by a summarizing comparison of them. Finally, section 5 concludes the paper and discuss areas of future work.

2. Concept Extraction

Concept extraction is very useful in many applications, such as search, classification, clustering, and for ontology learning from text. Extraction of domain specific concept is a key component in ontology construction from text. Manual concept extraction is time-consuming process. (Crangle et al.2004) defined concept extraction as follows: **"**Concept extraction is the process of deriving terms from natural-language text that are considered representative of what the text is about. The terms are natural-language words and phrases which may or may not themselves appear in the original text".

Concept formation should provide an intension definition of concepts, their extension and the lexical that are used to refer to them (Cimiano, 2006). Also, (Buitelaart al., 2005) considered that a concept should have a linguistic realization. Therefore, in order to identify the set of concepts of a domain, it is necessary to analyze a document to identify the important domain

terms that represent concepts, which can be a single word or multiword term. The importance of term is measured by modeling statistical features and linguistic features. The terms above a certain threshold are referred to concepts. Therefore, the major challenge in concept extraction is to be able to differentiate domain terms from non-domain terms (Zouaq and Nkambou 2011).

Many concept extraction methods have been proposed in the literature.TF-IDF (Term Frequency-Inverse Document Frequency) is a popular method that is widely used in information retrieval and machine learning fields. This method adopted by Text-To-Onto (Maedche and Staab, 2001), first, it employs a set of pre-defined linguistic filters (particularly the POS tag based rules) to extract possible candidate terms, including single-word terms and multi-word terms, from texts. Then, some statistical measures are used to remove irrelevant concepts.

Clustering techniques can be used to induce concepts. Based on Harris distributional hypothesis (Harris 1970), which stated that words that occur in similar contexts often share related meaning, the concept is considered as a cluster of related and similar terms. Also, Formal concept analysis and Latent semantic indexing algorithm used to build attribute-values pairs that correspond to concepts (Rizoiu and Velcin 2011). Another approach utilized WordNet to extract synonyms and relevant information about a given concept that contributes to concept definition (Zouaq and Nkambou 2010).

3. Terminology Extraction: Preliminaries and Definition

Terminology extraction is the first task on ontology construction from text. Its purpose is "to obtain from a domain corpus the most significant set of terms, that is, the set of superficial representations of domain concepts that better represents the domain for a human expert" (Pazienza et al., 2005). Terminology is the principle link between text and an ontology, which aims to map concepts to terms. A term is defined as a textual realization of a specific concept. Properties to define terms as termhood and unithood have been proposed in the literature. The termhood express the extent to which a linguistic unit is related to domain specific concepts, while the unithood express the degree of stability of syntagmatic collocations. The mapping of a term to a concept in an ontology is non-trivial. Since there is no one-to-one correspondence between concepts and terms. Two problems may appear in this case (Spasic et al., 2005, Astrakhantsev and Turdakov 2013):

- synonymy: occurs when several terms have the same concept as a denotation (Term variation);
- homonymy and polysemy: occurs when the same term refers to multiple concepts (Term ambiguity).

A classification of term variants for Arabic language can be found in Boulaknadel et al., 2008; Bounhas and Slimani 2009, Attia, 2010). They demonstrated the need for term variants recognition in the task of Arabic terms extraction. However, the majority of studies addressing this problem do not consider terminological variance. In the literature, there are many approaches, techniques and algorithms used for term extraction. The main methods to extract term fall into statistical methods, linguistic methods or hybrid ones. Researches interested on Arabic terms extraction are recent and most current works are applying the hybrid approach. The following sub sections provide a detailed description of the existing tools and methods for Arabic term extraction.

4. Terminology Extraction Approaches

The literature provides several ways to classify term extraction methods. It can be divided into two categories, based on the learning paradigm they employ:

- The approaches that extract keyphrase based on a supervised learning technique, which are regarded as intelligent way to summarize documents (El-shishtawy and Al-sammak, 2012). While this approach provides less noisy keyphrases, it needs many learning examples for machine.
- The approaches that extract keyphrase from documents, which is unsupervised learning technique, trying to discover concepts, rather than learn from examples. (El-beltagy and Rafea 2009) presented keyphrase system called KP-Miner, which extracts keyphrase candidates using some features: TF-IDF, position of phrase and boosting factor. Sine this approach does not depend on the expert, it scales well to large datasets. The major drawback is the large number of extracted phrases, most of them not interesting to a specific domain, that leading to a noisy output.

Other researchers (Buitelaar et al., 2005, Cimiano et al., 2006) divided term extraction methods according to their employed methods; Linguistic, statistical and hybrid.

4.1 Linguistic Approach

The linguistic approaches consider terms as candidate concepts and identify terms by using morphological and syntactic information about terms. Linguistic approaches perform a linguistic analysis to a text in order to obtain linguistic knowledge that will be used in term extraction and for subsequent phase of ontology learning from text (Al-Arfaj and Al-Salman 2015b).

Linguistic approaches utilize two classes of extraction methods:

- Based on Part Of Speeches (POS tagging), also referred as shallow text processing,
- Based on text structure dependencies (parser), also known as deep text processing.

According to literature, concepts are usually described by noun phrases, since noun phrases usually contain domain relevant semantic information. So, most of the linguistic and hybrid approaches focus on noun phrase extraction. The linguistic methods are used to extract noun phrases that constitute multiword terms. Multiword terms of a given domain belong to a finite set of syntactic structure. These patterns can be identified by an expert. For example, in case of medicine field from hadith corpus in Arabic language, the following patterns can be detected (Table 1).

Table1: Examples of Linguistics Patterns of noun Phrases for Ontological Terms extraction

Pattern	**Example**
Noun	الشفاء/ Treatment
Noun Adjective	الحبَّة السوداء/black cumin
Noun Preposition Noun	الحجامة من الداء/ To be cupped for a disease.

Some works used a pure linguistic method to extract Arabic terms. For example, (Attia 2006) proposed a method for extracting Arabic Multiword Expressions (MWEs) based on manual lexicon of MWEs. He used the regular expression to identify candidate terms and presented some linguistic variations such as, morphological, lexical and syntactic variations. The weakness of this approach is the absence of statistical measures to rank candidate terms.

The linguistic information alone is not sufficient to extract terms. Many irrelevant terms would be considered as domain terms. The statistical information is used to determine which terms are significant in a particular domain.

4.2 Statistical approach

In the statistical approaches, all important domain terms are considered as domain concepts and require statistical measures to determine the importance of terms. It is based on the information about the frequency and distribution of words within domain or corpus. The most popular measures for statistical term extraction are:

TF-IDF, which is based on information retrieval algorithms (Salton and Buckley 1988), can be used to measure the importance of individual terms contributing to documents. It can be computed for a given term by multiplying its frequency in the current document term frequency (TF) with its inverse document frequency (IDF) a measure that yields large values for terms that appear only in very few documents of the given document collection. Words with high TF-IDF ranking are then selected as relevant terms. TF-IDF tends to produce single word terms. However, Arabic concept often consists of multiple terms. Multiword extraction is a two-phase process. First, collocations and terms that appear together are determined. A lexical pattern based approach can be used according to Arabic grammatical characteristics. Then, from this list, unique collocations are filtered out. To measure dependency between the two words in the binary collocation, many statistical measures have been proposed in the literature such as Mutual Information (MI) measure, the LikeLihood Ratio (LLR) and the Chi-square. For the Arabic language, experiments are performed on some of the most known measures to judge their ability to identify lexically associated words (Boulaknadel et al., 2008, Saif and Ab Aziz 2011).

Another approach use a reference corpus to extract domain important terms, since domain specific terms tend to occur more in specialized text of their domain than in general corpus. As proposed by (Ahmad et al., 1999),, terms can be identified by comparing a word's relative frequency in a given domain corpus to its relative frequency in a large corpus (reference corpus) covering many aspects of everyday language. Words that occur significantly more in domain corpus than in reference corpus should be extracted as relevant terms. The strength of the statistical approach is in independence from natural language and from domain knowledge, which makes it scale to different languages and domains. Nevertheless, this approach tends to provide high frequency terms but ignore low frequency terms, generating what is called silence. While linguistic approach succeeds in getting more precise results, but they do not scale well to new domain or large datasets (Jacquemin and Bourigault 2001). However, methods based on linguistic techniques need statistical measure to filter irrelevant terms.

4.3 Hybrid Approach

The majority of the current studies on Arabic terms extraction applying the hybrid approach. This approach first uses linguistic filters particularly POS tag based rules to extract candidate terms, including single word and multiword terms from text. Then some statistical measures are used to rank domain relevant terms and remove irrelevant terms.

C/NC-Value (Frantzi et al., 2000) is a domain-independent method, combining linguistic and statistical information for the extraction of multiword and nested terms. It enhances the frequency measure by taking into account the fact that terms can be nested into each other.

Further, the approach also incorporates information from context words, which are strong indicators of the termhood of the terms.

(AL-Katib et al, 2010) adopted C-Value combined with LLR statistical measure to extract Multi Word Term (MWT) from Arabic corpus. They concentrated on compound nouns as an important type of MWT and select bi-gram term. Their approach relied on two filters: linguistic Filter, to extract candidate terms by using patterns based on the POS tagger proposed by (Al-Taani et al., 2009). They considered the sequence of nouns, as well sequences of nouns that are connected by a preposition in the candidate MWTs extraction. To rank candidate MWT, they used the LLR measure for the unithood and C-Value measure for the termhood. The authors concluded that LLR method could be used efficiently as significance of association measure between the two words in the bigram with precision value equals to 94%.

Another method proposed by (Boulaknadel et al., 2008) for MWT extraction in Arabic for environment domain. They identified candidate terms by first, using POS tagger proposed by (Diab 2004) then applying a set of predefined linguistic filters to extract multiword terms. Second, four statistical measures which are LLR, FLR, MI and t-score are used for ranking MWT candidates. They considered term relevant to the environment domain if it has already been listed in existing terminology database. Their experiment showed that the LLR, FLR and t-score measures outperform the MI measure and LLR outperform other methods with precision value equals to 85%. The weakness of this approach is the lack of a morphological analysis.

(Bounhas and Slimani 2009) presented a hybrid approach to extract MWTs from Arabic corpus. They extracted candidate terms by using Arabic morphological analyzer (AraMorph) that has been developed by (Hajic et al. 2005) and POS tagger by (Diab 2004). To reduce the morphological ambiguity, they developed the Morpho-POS Matcher that integrate the AraMorph and POS tagger. They used sequence identifier to detect compound noun boundaries. In addition, they used both syntactic rules based on the POS and the morphological features to recognize compound nouns. On the statistical filter, their approach used only the LLR that computes the correlation between two terms. For the bigram candidates, they obtained the precision value equals to 93%, which outperform those obtained by (Boulkandel et al., 2008) that used the same corpus and evaluation method. However, these results only for bigram MWTs.

(Attia 2010) proposed three complementary approaches to extract Arabic Multiword Expressions (MWEs) from heterogeneous data resources. The first approach crosslingual correspondence asymmetries, which relied on the correspondence asymmetries between Arabic Wikipedia titles and titles in 21 different languages. The second approach translation-based extraction, which employed the automatic translation of MWEs from Princeton WordNet 3.0 into Arabic using Google Translate, and utilized different search engines to validate the output. The third corpus-based statistics, which applied lexical association measures Pointwise Mutual Information (PMI), Chi-square to detect collocations in a large unannotated corpus. They lemmatized the text to reduce inflectional forms using MADA (Habash et al., 2009).

(Saif and Ab Aziz 2011) proposed a hybrid method for extracting the noun compound from Arabic corpus. For the candidate identification, they used lemmatization and POS by (Al-Gahtani et al. 2009) in order to filter the candidates and determine the variations. To rank

candidate term, association measures LLR, Chi-square, MI, and enhanced mutual information (EMI) are computed for each candidate. From their experimental results, they concluded that the LLR is the best association measure that achieved highest precision value 92% in the n-best list with n=100.

(Al-Balushi and Ab Aziz 2014) further considered nested Arabic noun compound including bi-gram, tri-gram, 4-gram and 5-gram. The linguistic method consisting of stemming and POS tagging used to extract candidates, while the statistical association measures which are NC-value, PMI and LLR used to filter the candidates. The dataset is the same that has been used by (Saif and Ab Aziz 2011) which contains an online Arabic newspaper. The authors showed that NC-value obtained the best result compared to PMI and LLR in terms of extracting nested noun compounds with precision value 81%. However, they did not consider the sequences of nouns that are connected by a preposition in the candidate MWTs extraction step. Moreover, they used small dataset.

(Zaidi et al 2010) presented a hybrid approach for extracting collocations from Crescen Quranic Corpus. They first, analyzed the text with AraMorph, then simple terms were first extracted using TF-IDF measure. They obtained precision value 88%. For collocations extraction, the authors used rule based approach and MI to enhance and filter the obtained results, which improved the precision from 0.5 to 0.86.

(Mashaan Abed et al., 2013) extracted Arabic terminology from Islamic corpus. In the linguistic filter, they used POS tagger to extract candidate MWTs matching given syntactic patterns. While in the statistical filter, they applied TF-IDF to rank the single word terms candidate, and statistical measures (PMI, Kappa, Chi-square, T-test, Piatersky- Shapiro and Rank Aggregation) for ranking the MWTs candidates. From the experiments, the authors indicated the effectiveness of Rank Aggregation compared to others association measures with precision value 80% in the n-best list with n=100.
However, as reported by (El Mahdaouy et al., 2013) most of the previous studies have been evaluated on 100 best candidate MWTs and they deal with bi-grams only. Moreover, they rely on LRR or a combination of LRR and C-value and ignore contextual information in the ranking step.

More recently, (El Mahdaouy et al., 2013) considered contextual information and both termhood and unithood for association measures at the statistical filtering. To extract MWT candidates, they applied syntactic patterns on the output of the POS tagger developed by (Diab, 2009). The authors addressed MWT variants through a morphological analysis of the extracted MWTs based on light stemming. Then for candidates ranking, several statistical measures have been used including C-value, NCvalue, NTC-value and NLC-value. Their experimental results showed promising results for the NLC-value measure in term of precision for both bi-grams and tri-grams on an environment Arabic corpus. Table 2 gives a summarization and a comparison of the Arabic Term extraction methods.

5. Conclusion and Future Work

In this paper, we have presented an overview of automatic term extraction approaches for concept extraction from Arabic domain corpus. Based on the literature, it can be concluded that linguistic and statistical approaches have some weakness when they are used alone. On one hand, the statistical approach is unable to identify rare terms. It focuses on the statistic features of terms and ignores linguistic and semantic knowledge. On the other hand,

linguistic approach is language dependent and cannot scale well in large datasets. To avoid the weakness and leverage advantage of approaches, most of Arabic research applied a hybrid method to extract Arabic term.

Table2: Summary of the Arabic Term Extraction Methods
(The methods are categorized according to the type of extraction and filtering technique)

Research	**Extraction method**	**Filtering method**	**Evaluation**
Boulaknadel et al., 2008	Linguistic patterns POS tagging	LLR, FLR, MI and t-score	LLR outperform other methods with precision value equals to 85%
Bounhas and Slimani 2009	Morpho-POS Linguistic patterns Morphological features	LLR	Precision value equals to 93%
EL-Katib et al, 2010	Linguistic patterns POS tagging Stemming	C-Value + LLR	Precision value equals to 94%
Zaidi et al.,2010	AraMorph Jape rule using Gate	TF-IDF MI	Precision value 88% for simple term extraction, 86% for collocation.
Saif and Ab Aziz 2011	Linguistic patterns POS tagging Lemmatization	LLR, chi-square, MI, EMI	LLR is the best association measure with precision value 92%
Mashaan Abed et al., 2013	POS tagging Lemmatization	PMI, Kappa, CHI-squire, T-test, Piatersky-Shapiro and Rank Aggregation	Rank Aggregation the best association measure with precision value 80%
El Mahdaouy et al., 2013	Linguistic patterns POS tagging Light Stemming	C-value, NC value, LLR + C-value NTC-value and NLC-value	NLC-value measure outperformed others with precision value 82%
Al-Balushi and Ab Aziz 2014	POS tagging Stemming	NC-value, PMI and LLR	NC-value outperformed PMI and LLR with precision value 81%

As can be seen from the comparison (Table 2), it is clear that previous studies on Arabic term extraction mainly focused on two processes: candidate extraction and candidate filtering. This implies that these two processes are important for Arabic terms extraction. Furthermore, even though each of the methods might have different applications, the choice of features seems to be quite similar among the existing methods. Our analysis revealed that the majority of related works applied shallow linguistic analysis (POS tagging and stemming), the results can be enhanced by using more linguistic approaches such as parser.

Despite efforts to combine statistical measures to extract Arabic terms, most existing Arabic terminology extraction algorithms are unable to produce rare terms (silence). Another limitation with the existing method for Arabic term extraction is that it extracts terms that are general and not relevant to a specific domain. Domain specific knowledge resources should

be used to support term extraction methods. Due to the limitation of Arabic domain specific knowledge such as domain-specific corpora and ontologies, existing Arabic term extraction methods face challenges in extraction Arabic term from domain specific texts. Also, the precision of term extraction can be improved by resolving term variations and grouping of similar terms. Finally, the evaluation of the methods can be improved by using domain knowledge resources.

The Arabic term extraction is a complex task. The choice of approach depends on the type of data resources available and the application. It remains open work how to extract terms that are relevant to a specific domain. We need a new method that integrates domain knowledge resources and the characteristics of a document to extract the concepts, which are semantically relevant to the domain. In the further work, we will design, implement and evaluate a method for Arabic term extraction and to rank by relevance to the specific domain. This constitute concepts layer for learning ontology from Arabic documents.

6. References

Al-Arfaj, A. and Al-Salman, A. (2015a). Ontology Construction from Text: Challenges and Trends. International Journal of Artificial Intelligence and Expert Systems (IJAE), 6(2), pp.15-26

Al-Arfaj, A. and Al-Salman, A. (2015b). Arabic NLP Tools for Ontology Construction from Arabic Text: An Overview. In Proceeding of International Conference on Electrical and Information Technologies, ICEIT'15 March 25-27, 2015 Marrakech, Morocco

Al-Arfaj, A. and Al-Salman, A. (2014). Towards Ontology Construction from Arabic Texts- A Proposed Framework. In Proceeding of The 14th IEEE International Conference on Computer and Information Technology (CIT 2014), pp. 737-742

AL-Balushi, H and AB AZIZ, M. (2014). A hybrid Method of Linguistic Approach and statistical method for Nested Noun Compound extraction. Journal of Theoretical & Applied Information Technology, 67(3), pp. 601-608

Al-Gahtani S, and Black W and Mc-Naught J.(2009). Arabic part-of-speech-tagging using transformation-based learning. In Proceedings of the 2nd International Conference on Arabic Language Resources and Tools, Cairo, Egypt, The MEDAR Consortium, pp:66-70

Al-Tanni A and Abu-Al-Rub S. (2009). A rule-based approach for tagging nonvocalized Arabic words. The International Arab Journal of Information Technology, 6(3), pp.320-328.

Attia, M. (2006). An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks. In Challenges of arabic for NLP/MT conference, the british computer society, london, UK.

Attia, M., Toral, A., Tounsi, L., Pecina, P. (2010). Automatic Extraction of Arabic Multiword Expressions. The 7th Conference on Language Resources and Evaluation (LREC).

Ahmad, K., Gillam, L and Tostevin, L. (1999). Weirdness indexing for logical document extrapolation and retrieval (wilder). In the Eight Text Retrieval Conference (TREC-8).

Astrakhantsev, N and Turdakov, D. (2013). Automatic Construction and Enrichment of informal ontologies: A survey. Programming and Computer Software, 39(1), pp. 34-42

Buitelaar, P., Cimiano, P., Magnini, B.(2005). Ontology Learning from Text: An Overview. In Ontology learning from text: methods, evaluation and applications.

Breuker J, Dieng R, Guarino N, Mantaras RLd, Mizoguchi R, Musen M, editors. Amsterdam, Berlin, Oxford, Tokyo, Washington DC: IOS Press.

Beseiso, M., Ahmad, A and Ismail, R. (2010). A Survey of Arabic Language Support in Semantic Web. International Journal of Computer Applications. 9(1), 35-40.

Beseiso, M., Ahmad,A and Ismail,R. (2011). An Arabic language framework for semantic web. In proceeding of International Conference on Semantic Technology and Information Retrieval (STAIR).

Black, W., Elkateb, S.,Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A and C. Fellbaum. (2006). Introducing

the Arabic WordNet Project. In proceedings of the Third International WordNet Conference.

Boulaknadel, S., Daille, B and Aboutajdine, D. (2008). A multi-word term extraction program for Arabic language. In proceeding of the 6the international conference on Language Resources and Evaluation, Morocco, pp. 1485-1488.

Bounhas, I. and Slimani, Y. (2009). A hybrid approach for Arabic multi-word term extraction. In Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE), Dalian, China, pp. 429-436

Cimiano. P, Volker. J, and Studer. R. (2006). Ontologies on Demand? – A Description of the State-of-the-Art, Applications, Challenges and Trends for Ontology Learning from Text. Information, 57 (6-7), 315-320.

Cimiano, P. (2006). Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. In Studies in Philosophy and Religion, Springer.

Crangle, C., Zbyslaw, A., Cherry, M. and Hong, E. L. (2004). Concept Extraction and Synonymy Management for Biomedical Information Retrieval. In Proceedings of the Thirteenth Text REtrievel Conference (TREC 2004).

Diab, M. (2009). Second Generation AMIRA Tools for Arabic Processing: Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking. In proceeding of second International Conference on Arabic Language Resources and Tools. Eygpt, The MEDAR Consortium, pp. 285–288.

Diab, M., Hacioglu, K and Jurafsky, D. (2004). Automatic Tagging of Arabic Text: From raw text to Base Phrase Chunks. In proceedings of HLT-NAACL. pp. 149-152

El-Beltagy, S and Rafea A. (2008). KP-Miner: A Keyphrase Extraction System for English and Arabic Documents. Information systems. 34(1),132-144

El-shishtawy ,T and Al-sammak, A. (2012). Arabic Keyphrase Extraction using Linguistic knowledge and Machine Learning Techniques. In proceeding of the 2nd International Conference on Arabic Language Resources and Tools.

El-Mahdaouy, A., Said El Alaoui Ouatik and Gaussier, E. (2013). A study of association measures and their combination for Arabic MWT extraction. In Proceedings 10th International Conference on Terminology and Artificial Intelligence, pp. 45-52

Farghaly, A. & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. ACM Trans. Asian Lang. Inform. Process. 8, 4, Article 14), 22 pages.

Frantzi K., Ananiadou S and Mima H. (2000). Automatic Recognition of for Multi-word terms: the C-Value/NC-value method. International Journal of Digital Libraries, 3(2), pp. 117-132.

Harris Z. (1970). Distributional structure. structural and transformational linguistics, pp.775–794

Hajic, J., Smrz, O., Buckwalter, T and Jin, H. (2005). Feature-Based tagger of approximations of Functional Arabic morphology. In proceeding of the Fourth workshop on Treebanks and linguistic theories (TLT 2005), pp.53-64.

Jacquemin, C., and Bourigault, D. (2001). Term Extraction and Automatic Indexing. In R. Mitkov, editor, Handbook of Computational Linguistics. Oxford University Press, Oxford.

Maedche A and Staab S. (2001). Ontology Learning for the Semantic Web. IEEE Intelligent Systems, Special Issue on the Semantic Web, 16(2), 72 –79.

Mashaan Abed, A., Sabrina Tiun and AlBared, M. (2013). Arabic Term Extraction using Combined Approach on Islamic document. Journal of Theoretical & Applied Information Technology, 58 (3), pp.601-608

Nizar Habash,N., Rambow, O and Roth, R. (2009). MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, The MEDAR Consortium.

Pazienza, M., Pennacchiotti, M., and Zanzotto, F. (2005). Terminology Extraction: An Analysis of Linguistic and Statistical Approaches, Knowledge Mining, ser.: Studies in Fuzziness and Soft Computing, Sirmakessis, S., Ed., Berlin/Heidelberg: Springer, vol. 185, pp. 255– 279.

Rizoiu, M and Velcin, J. (2011). Topic Extraction for Ontology Learning. Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances, Wong W., Liu W. and Bennamoun M. eds. (Ed.). pp. 38-61.

Saif, A and Ab Aziz, M. (2011). An Automatic Collocation Extraction from Arabic Corpus. Journal of Computer Science. 7 (1), 6-11.

Salton, G and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24 (5), 513–523

Spasic I., Ananiadou S., McNaught J and Kumar A. (2005). Text mining and ontologies in biomedicine: Making sense of raw text. Bioinformatics, 6(3),pp. 239-251

Zaidi, S., Laskri, M and Abdelali, A. (2010). Arabic collocations extraction using Gate. In Proceeding international conference on Machine and Web Intelligence (ICMWI), pp. 473 - 475.

Zouaq. A and Nkambou, R. (2010). A Survey of Domain Ontology Engineering: Methods and Tools. In Bourdeau & Mizoguchi (Eds): 'Advances in Intelligent Tutoring Systems', Springer, pp.1-20.

Zouaq, A., Gasevic, D and Hatala, M. (2011). Towards open ontology learning and filtering. Information Systems, 36(7), 1064–1081.