A New Concept Extraction Method for Ontology Construction From Arabic Text

Abeer Alarfaj *aaalarfaj@pnu.edu.sa*
Department of Computer Sciences
College of Computer and Information Sciences
Princess Nora Bint AbdulRahman University
Riyadh, Saudi Arabia

Abdulmalik Alsalamn *salman@ksu.edu.sa*
Department of Computer Science
College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia

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. In this paper, we present a new concept extraction method for Arabic ontology construction, which is the part of our ontology construction framework. A new method to extract domain relevant single and multi-word concepts in the domain has been proposed, implemented and evaluated. Our method combines linguistic, statistical information and domain knowledge. It first uses linguistic patterns based on POS tags to extract concept candidates, and then stop words filter is implemented to filter unwanted strings. To determine relevance of these candidates within the domain, different statistical measures and new domain relevance measure are implemented for first time for Arabic language. To enhance the performance of concept extraction, a domain knowledge will be integrated into the module. The concepts scores are calculated according to their statistical values and domain knowledge values. In order to evaluate the performance of the method, precision scores were calculated. The results show the high effectiveness of the proposed approach to extract concepts for Arabic ontology construction.

Keywords: Ontology Construction, Arabic Ontology, Arabic Language Processing, Concept Extraction, Arabic Term Extraction, Specific Domain Corpus.

1. INTRODUCTION

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 [1]. 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 [9].

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. All these difficulties pose a significant challenge to researchers developing NLP systems in general and particularly on the terminologies extraction for Arabic [1].

To build Arabic ontology, the first step is to find the important concepts of the domain. The concept linguistically 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 [11]; Boulaknadel et al [5]; Bounhas and Slimani [6]; Saif and AbAziz[18]). However, key phrase extraction is different from concept extraction. In the framework by (Al-Arfaj and Al-Salman [3]), 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.

According to the study [1], it remains open work how to extract and rank single and multi-word concepts by relevance to the domain.

In this paper, we first describe the new proposed method for concept extraction from Arabic text. Then we present experiments used to evaluate its performance with medicine documents from hadith corpus.

The distinctive features of our method for concept extraction are as follows:

- Our method is unsupervised; it does not need training data.
- Also, our method does not use contrasting corpora to identify domain concepts. This leads to avoiding the problem of skewness in terms frequency information.
- It can extract domain relevance concept using a combination of statistical measures and domain knowledge. This overcomes the problems that have been found in the methods that are based only on the frequency or TF-IDF to measure the importance of the candidates.

The main contributions of the paper are as the following:

- Propose a new method for concept extraction from Arabic text.
- Compare different statistical algorithms for term weighting and extraction.

The rest of the paper is organized as follows. A brief overview of the existing works in concepts extraction is summarized in section 2. Section 3 provides the details of the proposed method and its main steps. Section 3.1 describes the pre-processing steps. The algorithm used and implemented in this study for candidate concept extraction is described in Section 3.2. Section 3.3 presents the concept extraction selection algorithm. In Section 3.4, a candidate concepts selection algorithm using domain knowledge is presented. Section 3.5 describes the post-processing techniques used in this study. Section 4 provides a comparative evaluation of the method in terms of precision, summarizes the finding and results. Finally, section 5 concludes the paper and discuss areas of future work.

2. RELATED WORKS

The main objective of concept extraction task is to identify domain concepts from pre-processed documents for the domain being investigated. Concept extraction is a primary and the basic layer for ontology learning from text and very useful in many applications, such as search, classification, clustering. The key challenge is how to extract domain specific concepts that represent the key information of a corpus in a domain of interest.

Concept formation should provide an intension definition of concepts, their extension and the lexical that are used to refer to them [9]. Also, [7] 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 [4].

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 [8], 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 [24]), 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 [23]. Another approach utilized WordNet to extract synonyms and relevant information about a given concept that contributes to concept definition [4].

In [20] unsupervised approach used for concept extraction from clinical text. Their method combined natural language descriptions of concepts with word representations, and composing these into higher-order concept vectors. These concept vectors are then used to assign labels to candidate phrases which are extracted using a syntactic chunker. They reported an exact F-score of.32 and an inexact F-score of.45 on the well-known I2b2-2010 challenge corpus.

Authors in [17] proposed a method for concepts extraction using semantic information, semantic graph-based Chinese domain concept extraction (SGCCE). First, the similarities between terms are calculated using the word co-occurrence, the LDA topic model and Word2Vec. Then, a semantic graph of terms is constructed based on the similarities between the terms. Finally, according to the semantic graph of the terms, community detection algorithms are used to divide the terms into different concept clusters. The experimental results showed the effectiveness of their proposed method. The results of the concept extraction are used for relation extraction tasks [16].

A method for MWT extraction in Arabic for environment domain is proposed in [5]. The authors identified candidate terms by first, using linguistic filters. Second, four statistical measures which are LLR, FLR, MI and t-score are used for ranking MWT candidates. 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%.

[22] Presented a hybrid approach for extracting collocations from Crescen Quranic Corpus. They first, analyzed the text with AraMorph, and 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 they reached precision value 0.86%.

In [26] authors extracted Arabic terminology from Islamic corpus. They used the linguistic filter to extract candidate MWTs matching given syntactic patterns. 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.

The method in [27] considered contextual information and both termhood and unithood for association measures at the statistical filtering. To extract MWT candidates, the authors applied syntactic patterns and for candidates ranking, several statistical measures have been used including C-value, NCvalue, NTC-value and NLC-value. Their experimental results showed that

NLC-value measure outperformed others with precision value 82% on an environment Arabic corpus.

An extensive analysis of term extraction approaches and the summary of Arabic terminology extraction methods, with main intent of highlighting their strengths and weaknesses on extract domain relevant terms detailed in [1].

In this work, we implemented the five algorithms for concept extraction, Concept Frequency-Document Frequency (CF-DF), Term Frequency-Inverse Document Frequency (TF-IDF), Concept Frequency-Inverse Document Frequency (CF-IDF), Relative Concept Frequency (RCF), Average Concept Frequency in the corpus (Avg-CF).

We investigate the performance of these algorithms and demonstrate that CF-DF algorithm is the best choice for concept extraction. We also discuss why one method performs better than other and what could be done to further improve the performance.

Our contributions to the Arabic concept extraction field are as follows. We evaluate and compare different statistical measures and proposed a new one for candidates' concepts weighting. Our method can extract rare concepts, even those appearing with low frequency. It also excludes irrelevant concepts even if they occur frequently in the corpus.

We assessed the effectiveness of our method by computing the precision for each experiment. For evaluation purposes, we focus on the medicine domain from hadith corpus. We observe that our method for concept extraction from Arabic text significantly improves precision.
The output list from this module constitutes the fundamental layer of ontologies.

3. PROPOSED METHOD FOR ARABIC CONCEPT EXTRACTION

This section describes the baseline measures and the new method that we propose for Arabic concept extraction and ranking based on linguistic, statistical information and domain knowledge. Our method for concept extraction has four main steps; described in Figure 1: pre-processing, linguistic filter to extract candidate terms (single and multi-word), candidate selection according to different weighting models, and post-processing (refinement and expert validation).

After the preprocessing phase, the annotated corpus will be input to concept extraction phase. In this phase, terms representing concepts will be extracted. First, we apply linguistic filter (Pattern on POS annotated text) to extract concept candidates. Additionally, a stop word will be used to eliminate terms that are not expected to be a concept in the domain, which improves precision of the output candidate terms. Second, we calculate the weights of candidates using a combination of statistical and domain knowledge. The higher the weight of a candidate is, the more relevance to the domain. Statistical information is obtained using different statistical measures (CF-DF, TF-IDF,CF-IDF, RCF, Avg-CF). Third, since all the statistical measures are based on the frequency only, the multiword terms and the concepts, which are important to the domain but has less frequency, will not be extracted. To enhance the performance of concept extraction, a domain knowledge will be integrated into the module. The domain knowledge is obtained from a domain specific list extracted from the taxonomy of the corpus.

Finally, our method generates a ranked list of key concepts according to their weights. Average threshold is computed to prune incorrect concepts. The concepts will be displayed to the expert to choose the valid concept and to add the missing one. In the following, we describe each of these steps in more details.

3.1 Pre-Processing

Arabic documents were processed in the steps described in [2]. After the preprocessing phase, the annotated corpus will be input to concept extraction phase. More details of preprocessing steps and tools are presented in [2]. First preprocessing phase is implemented. Table1 shows the sample hadith text before and after preprocessing steps.

Normalization

Many level of orthographic normalization is carried on input documents before analyzing the text. This includes normalization of the hamzat alif to a bar alif, normalization of taa marbutah to haa, and removal of extra spaces between words. In this paper we work on the texts without diacritics, since these diacritics have no effect in determining concepts and relations between them and generate a problem and overhead in most of the Arabic morphological analyzer. In this work the Ambiguity is solved by using context, for example, the word العين Has context {حق العين النظره, العين }and {العين ,الكحل, امراه}, so, according to the context we can determine the mean of the words.

These steps performed for normalizing the input text.

- Convert to UTF-8 encoding
- Remove diacritics and tatweel.
- Remove non-Arabic letters.
- Replace (إ آ أ) by (ا)
- Replace (ة) by (ه)
- Remove extra spaces between words
- Separate punctuations from words

Sentence Segmentation and Tokenization

The output of this phase is individual sentences to be considered for further processing. Each of the individual sentences is given as input to the next module of parsing. After sentence segmented, the next step is to provide the sentence to the Stanford POS tagger for tagging the tokens.

3.2 Candidate Concept Extraction

The main objective of this step is to extract all possible candidates for concepts using linguistic techniques. The details steps are as follows.

3.2.1 The Linguistic Filter

We use the Stanford POS tagger for Arabic text [19] to perform the linguistic analysis phase. The tagger assigns for each token in the sentence its grammatical category. For example, in Table 1, each sentence in the hadith text, the tagger identifies nouns, verb, and preposition. We chose the Stanford POS tagger because it reaches an accuracy of a 96.42% on Arabic and it is written in java that can be easily integrated in our module.

Arabic terms consist mostly of nouns and adjectives and sometimes prepositions. We observed that many of the terms mentioned in the hadith corpus are multi-word terms. The syntactic structure of multi-word terms can be used in semantic relations extraction between concepts.

Preprocessing steps	**Sample hadith Text**
Sample hadith Text	**الشفاء في ثلاثة: شربة عسل، وشرطة محجم، وكيّة نار، وأنهى أمتي عن الكي.**
Text after normalization	**الشفاء في ثلاثه : شربه عسل ، وشرطه محجم ، وكيه نار ، وانهى امتي عن الكي.**
Text after sentence segmentation	**الشفاء في ثلاثه** **شربه عسل** **وشرطه محجم** **وكيه نار** **وانهى امتي عن الكي**
Text after Stanford POS tagger	الشفاء DTNN/في IN/ثلاثه CD/ شربه NN/عسل NN/ و CC/شرطه NN/محجم JJ/ و CC/كيه NN/نار NN/ و CC/انهى VBD/امتي NNP/عن IN/الكي DTNN/
The list of Noun phrases after using linguistic filters	**الشفاء , شربه عسل, شرطه, شرطه محجم , كيه نار , امتي, الكي**
The list of Noun phrases after stemming	**شفاء, شرب عسل, شرط,شرط محجم,كيه نار, امتي , الكي**
The list of Noun phrases after root extraction	**شفو,شرب عسل, شرط ,شرط حجم, كيه نير, امتي , الكي**
Text after MADA+Stanford POS tagger	الشفاء DTNN/في IN/ثلاث NN/ه PRP$/ :PUNC/شرب VBD/ه PRP/عسل NN/ ، JJ/و CC/شرط NN/ه PRP$/محجم NN/ ،NNP/و CC/كي NNP/ه PRP$/نار NN/ ،JJ/و CC/انهي VBD/امة NN/ي PRP$/عن IN/الكي DTNNP/ ./PUNC
The list of Noun phrases after using linguistic filters	**الشفاء, ثلاث,عسل,شرط,محجم, كي,نار,امة, الكي**

TABLE 1: A sample hadith text from Medicine Book after linguistic analysis steps.

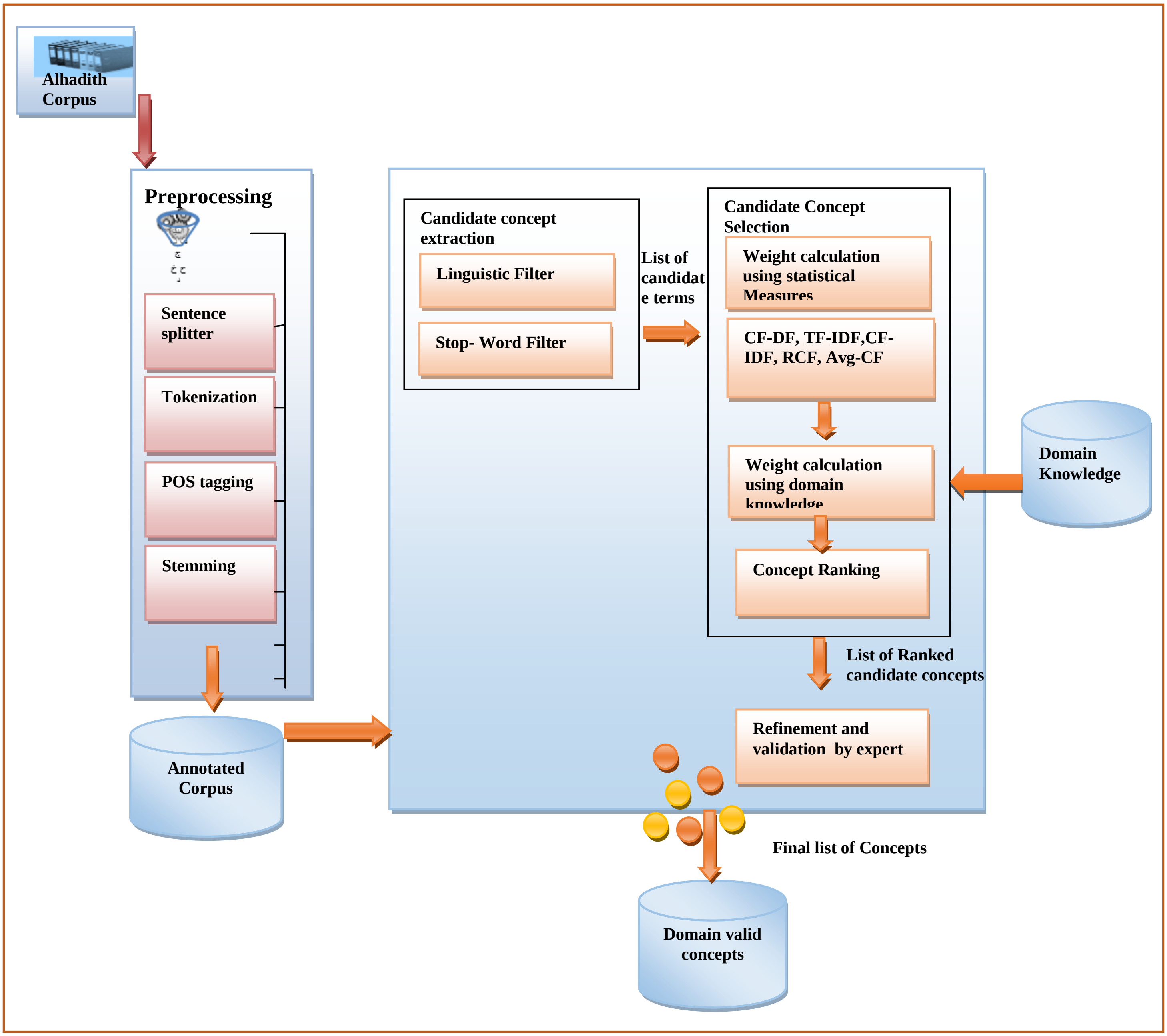
Alhadith Corpus
Preprocessing
Sentence splitter
Tokenization
POS tagging
Stemming
Annotated Corpus
Candidate concept extraction
Linguistic Filter
Stop- Word Filter
List of candidat e terms
Candidate Concept Selection
Weight calculation using statistical Measures
CF-DF, TF-IDF,CF-IDF, RCF, Avg-CF
Weight calculation using domain knowledge
Concept Ranking
Domain Knowledge
List of Ranked candidate concepts
Refinement and validation by expert
Final list of Concepts
Domain valid concepts

FIGURE 1: The Proposed Concept Extraction Method.

3.2.2 Candidate Concepts Extraction Following Patterns

Concepts in ontology represent a set of classes of entities or things within a domain [14]. According to literature, concepts are usually described by noun phrases, so we considered every noun phrase in the document as a candidate term. Therefore, for extracting noun phrase candidates we implemented linguistic filters based on predefined POS patterns to extract noun phrases that constitute multi-word terms.

First, we applied Stanford POS tagger prior to linguistic filters. Next step is to determine patterns for noun phrase concepts using tags. The algorithm for extracting noun phrases based on the patterns shown in Algorithm 1. Noun phrase consist of one head noun followed by one or more nouns or adjective [5,6]. A linguistic filter is applied on each tagged sentence to extract candidate multiword terms. Based on our inspection of the concepts we identified by studying medicine book from hadith corpus, we implemented all the filters in Table 2 for our experiments. In Table 1, we see the example of the noun phrases extracted from the sample hadith text. Many candidate terms identified by this method are not key concept. Therefore, several filters are used to reduce the number of candidate terms. The first filter is stop words.

Pattern	Example
(Noun)+	الشفاء
(Noun)+ (Adjective)+	الحبّة السوداء
Noun Preposition Noun	الحجامة من الداء

TABLE 2: Examples of Linguistic patterns of noun phrases for ontological terms extraction.

3.2.3 Stop Word Filter

We notice that the POS tagger returns the terms such as شيء, احدهما, احيانا as nouns. In order to improve precision of the output candidate terms, we have implemented our own stop words filter to eliminate terms that are not expected to be a concept in the domain. The list of stop words is constructed based on the domain observations. Further, stop words are not allowed at the beginning or end of the phrase. If a phrase starts or ends with a stop word, it will be removed from the list of candidate terms and keep ones in the extracted noun phrase for relation extraction. (examples: "الحجامه من الداء" "الحجامه على الراس").

In Table 3, we see the example of the candidate terms after applying stop word filter.

The list of Noun phrases after linguistic filter	The list of Candidate terms after stop word filter
العسل,الحلواء,وسلم,عليه,الله,النبي	النبي , الله, الحلواء,العسل

TABLE 3: Example of using Stop Word Filter.

```
Algorithm: Linguistic Filters
Input: Set of statements S in Preprocessed Text
Patterns:
Pattern 1= If a noun is followed by one or more noun
Pattern 2= If one or more noun is followed by one or more adjective
Pattern 3= if a noun is followed by preposition followed by a noun
Output: set of noun phrases NPlist
Procedure:
Begin
  For each statement Si in S do
    Ci ← Stanford POS tagger (Si)
    For each Pattern in Patterns
    NPi = Apply Pattern(Ci)
    Add NPi to NPlist
    i←i+1
End
 For each NPi in NPlist
   For each word in NPi
     If NPi[head]or NPi[end] exists in Stop word list
        Remove NPi  from NPlist
   End
 End
return NPlist
End
```

ALGORITHM 1: Candidate Concepts Extraction Algorithm.

3.3 Candidate Concept Selection

After extracting candidate terms using the linguistic filter, our method assigns weights to these candidate terms, ranks them, and extracts the ones with higher weights. The weight of concept candidate is determined using a combination of the statistical and domain knowledge. It is based on the following idea: candidate term that is included in the domain specific list is likely to be a domain concept. Statistical knowledge is obtained from the different statistical measures implemented in this module. To measure the relative importance of the concept candidates in the specific domain, we use a domain list that consists of domain key concepts. The list of concept candidates is parsed to determine if it is found in domain specific list.

Our method assigns higher weights to more domain specific concepts; candidates that are frequently appear in the corpus and in the domain specific list. All the weights of concept candidates are normalized to range from 0 to 1 after they are calculated. All candidate concepts are then ranked in descending order by their weights.

The domain concepts can be extracted from the ranked list according to different parameters either by defining a specific number of concepts to be extracted or by setting the average frequency as threshold for concepts to be extracted. Algorithm 2 shows the algorithm for candidate concepts selection and ranking using different statistical measures.

Details of the proposed method for candidate concepts selection are presented in the following sub-sections.

```
Algorithm: Candidate concepts selection
Purpose: extract candidate concepts and ranking using statistical measures
Input: corpus= set of documents of a specific domain
Avg-threshold= frequency threshold for candidate terms
Output: Lterms: list of ranked terms
Begin
Read corpus
Normalizing input text file
Sentence segmentation
Tag the tokens
Extract candidate terms t by filtering with patterns
Apply stop word filter for each candidate term t
For each candidate term t Apply statistical measures
  CF-DF-value(t)= CF(t)xDF(t)
  Add t to Lterms
End
Rank Lterms by the value obtained CF-DF-value
Compute Avg-threshold
Select from Lterms the candidate above Avg-threshold
End
```

AGORITHM 2: The Candidate Concepts Selection and Ranking.

3.3.1 Candidates Concepts Selection using Different Statistical Measures

Although our linguistic filter returns noun phrases, it may include phrases which do not belong to domain-specific (e.g. 'امتي' in Table 1). In order to refine the results, weighting models which are the statistical measures to determine the importance of terms has been employed which places highest weighted terms as the most important concepts. We have implemented different algorithms to assign weights for every candidate terms.

3.3.2 Weight Calculation of Concept Candidates

After multi-word terms extraction using linguistic filters, it still suffers from the problem that these extracted phrases may not cohesive enough to be treated as a term. For tackling this problem [6,22] use statistical measures such as Mutual Information measure and LikeLihood ratio test to score the extracted units. However, term frequency has shown to produce better performance than other measures for multi-word term extraction [10].

Our goal is to extract domain specific terms. To measure the importance of concept candidates in the corpus of the domain, there are different algorithms based on the following idea: terms which occur most frequently are considered as the relevant terms of the domain.

Text2Onto [8] extracts concepts using different weighting schema (entropy, Relative Term Frequency (RTF), TF-IDF) to measure their domain relevance. In the OntoLearn system [13], to extract domain concept two statistical measures Domain Relevance (DR), Domain Consensus (DC) are used together. However, this measure doesn't consider the rare concepts. Also, the domain relevance measure depends on the term's frequency in the target domain corpus and the contrastive corpus.

(Jiang & Tan [12] extract concept by first, noun phrase extraction from the corpus using linguistic filter. Then, they compute the frequency of these phrases in the target corpus and contrastive corpus to assign higher weight to more domain specific term.

The problems of using contrastive corpus are: If there is a large difference between the size of corpus and the contrastive corpora, it will leading to high skewness in the frequency of noun phrases among the domain. For tackling this problem, in our algorithm to measure domain relevance we rely on [12,13] measures and we did not use contrastive corpus to measure importance of concepts to the target domain.

The assignment of weights is made using one of the following algorithms:

Concept Frequency-Document Frequency (CF-DF)

The base line TF-IDF measures the term relevancy to a given document in the document collections. In this thesis, we proposed a new algorithm for extracting conceptual terms from domain texts. The proposed method is based on the assumption that document frequency of a concept in the document collection is a good measure for estimating concept relevance to specific domain. The idea is to measure concept relevance to the domain, if a concept occurs in multiple documents, it is considered more relevant compared to with those occurring only in single document. Each concept obtains the value from its frequency in the document multiplied by the frequency of the same concept in the documents collection.

Let C be the set of candidate concepts extracted using the linguistic filter. The weight of a candidate $c \in C$ is computed as:

$$CF - DF(c,d) = \frac{f(\mathrm{c,d})}{\mathrm{Max}\ f(\mathrm{c,d})} X \frac{\mathrm{D}f(\mathrm{c})}{\mathrm{Max\ D}f(\mathrm{c})} \qquad (1)$$

Where Concept Frequency (CF) is the number of occurrences of each concept c in the document d, and then normalized to prevent bias towards longer document by dividing the count by the max concept frequency in the document d for all candidates $c \in C$.

Document Frequency (DF) is the number of documents in a corpus that contain the concept c.

From the equation 1, we can see that, the higher the frequency of concepts the more likely the concept is to be important.

Term Frequency-Inverse Document Frequency (TF-IDF)

Term Frequency TF(t,d) is the number of times term t occurs in the document d. That measures popularity of term in document, which is normalized with the maximum term frequency in the document as shown in equation 2. TF is usually normalized to prevent a bias towards longer documents which may have a higher term count regardless of the actual importance of that term in the document.

$$TF(t,d) = \frac{f(\mathrm{t,d})}{\mathrm{Max}\ f(\mathrm{t,d})} \qquad (2)$$

Where:
$f(\mathrm{t,d})$ is the frequency of occurrence of term t in the target document d.
$\mathrm{Max}\ f(\mathrm{t,d})$ is the max frequency in the document d for all t.

Document Frequency (DF) is the number of documents in a corpus that contain the term.

Inverse Document Frequency (IDF) is the measure of the importance of the term to the corpus in general terms. Estimate the rarity of a term in the whole document collection; attempt to automatically remove terms that are not important because they are common on the whole corpus. It means that if a term occurs in all the documents of the collection, its IDF is zero.

This is computed by dividing the number of documents in the corpus by the number of documents that contain that term (DF) and then taking the logarithm of that quotient.

Using the following formula:

$$IDF(t) = \log \frac{|\mathrm{D}|}{|\{\mathrm{d}: \mathrm{t} \in \mathrm{d}\}|} \qquad (3)$$

Where:
$IDF(t)$ is the inverse document frequency for term t
$|\mathrm{D}|$ is the total number of documents
$\{\mathrm{d}: \mathrm{t} \in \mathrm{d}\}$ is the number of documents containing t

Given that if term t does not occur in the corpus, the denominator can leads t division by zero. So it is common to add 1 as shown in equation 4.

$$IDF(t) = \log \frac{|\mathrm{D}|}{1 + |\{\mathrm{d}: \mathrm{t} \in \mathrm{d}\}|} \qquad (4)$$

For each candidate concepts, we calculate TF-IDF. To achieve this, equation 2 and 4 are combined to form equation 5.

$$TF - IDF(t,d) = TF(t,d)x\ IDF(t) \qquad (5)$$

Finally, normalization is done by dividing TF-IDF value of all concepts by square root of the sum of square of TF-IDF value of each concept.

The terms with high values of TF-IDF are important terms, since they have a high TF in the given document and low occurrences in the remaining documents in the corpus which filtering out common terms. TF-IDF tends to extracting more single words terms than multi word terms concepts. This is because multi word terms are less likely to appear than single word terms. While in domain specific corpus, the domain concepts are multi word terms and the probability of concepts occurring in many documents is high.

Also, TF-IDF assigning higher weights to rarer candidate concepts; if the important domain oncepts appear in most of domain documents it will be discarded since its IDF value tends to be zero. Thus, the TF-IDF may perform poorly in some context. However, in our work we observed that the important concepts in the domain are not appearing frequently in most of the domain documents.

Concept Frequency-Inverse Document Frequency (CF-IDF)
CF-IDF is a modified version of TF-IDF to handle multi-word terms [25]. For each candidate concepts, the weight is calculated by the following:

If concept length =1 then

$$CF - IDF(t,d) = CF(t,d)x\ IDF(D) \quad (6)$$

Else

$$CF - IDF(t,d) = \ CF(t,d)\ x\log(D) \quad (7)$$

Where:
$CF(t,d)$ is the number of times a term occurs in a document d.
D is the total number of documents in the corpus.

For single word concept the weight is computed using equation 6, where $IDF(D) = \frac{|D|}{DF}$ is computed from equation 4. For multi word concept, the weight is computed using equation 7, which set DF to 1. This is due to that multi word terms do not occur frequently within a collection of documents as single word terms.

Relative Concept Frequency (RCF)
It calculates relative term frequency which is calculated by dividing the absolute concept frequency of the concept c in the document d (number of times concept c occurs in document d) divided by the maximum absolute concept frequency of the document (the number of times any concept occurs the maximum number of times in the document d). It is computed in the following way:

$$RCF(c,d) = \frac{|f(c,\mathrm{d})|}{Max\ |f(c,d)|} \quad (8)$$

Average Concept Frequency In Corpus (Avg-CF)
Avg-CF is calculated by dividing the total frequency of a concept in a corpus by its document frequency [21]. For each document in the corpus *D*, we aggregate concept frequencies across all documents *D*. Formally, given a candidate *c* ∊ *C*, its weight with respect to *D* is calculated as follows:

$$Avg - Cf(c,d) = \sum_{i=1}^{|D|} f(c,di) \backslash f(c,d) \quad (9)$$

Where $f(c,di)$ is frequency of the concept c in the document d for all $d_i \in D$.

The ranked list of extracted candidate concepts, shown in Table 4, illustrates how our methods select the most relevant concept to the domain.

CF-DF	TF-IDF	CF-IDF	RCF	Avg-CF
ماؤها شفاء للعين: 1.0	وريقه بعضنا: 0.090746	وريقه بعضنا: 0.100586	شفاء: 1.0	الله: 1.0
لكل: 1.0	وجهها سفعه: 0.090746	وجهها سفعه: 0.100586	وريقه بعضنا: 1.0	النبي: 0.247422
شهيد: 1.0	هوامك: 0.090746	نار جهنم: 0.100586	وجهها سفعه: 1.0	الم: 0.221649

شرطه: 1.0	نسيكه: 0.090746	مهر البغي: 0.100586	هوامك: 1.0	رسول الله: 0.149484
سحر: 1.0	نار جهنم: 0.090746	ماؤها شفاء للعين: 0.100586	نسيكه: 1.0	ناس: 0.103092
رب الناس: 1.0	مهر البغي: 0.090746	كراهيه المريض: 0.100586	نار جهنم: 1.0	شفاء: 0.082474
بارض: 1.0	ماؤها شفاء للعين: 0.090746	فيح جهنم: 0.100586	مهر البغي: 1.0	نار: 0.056701
الوشم: 1.0	لاعدوى: 0.090746	فوح جهنم: 0.100586	عجوه: 1.0	داء: 0.056701
المن: 1.0	كراهيه المريض: 0.090746	عدوى ولا طيره: 0.100586	المجذوم: 1.0	عين: 0.051546
الله: 1.0	فوح جهنم: 0.090746	سبع تمرات: 0.100586	ماؤها شفاء للعين: 1.0	لكل: 0.046391
الكماه: 1.0	عينها: 0.090746	رب الناس: 0.100586	لكل: 1.0	عدوى: 0.046391
العين حق: 1.0	عدوى ولا طيره: 0.090746	ذي حمه: 0.100586	لاعدوى: 1.0	كلمه: 0.0412371
الحمى: 1.0	شهاده: 0.090746	ثمن الكلب: 0.100586	كراهيه المريض: 1.0	طيره: 0.041237
الحجام اجره: 1.0	سته: 0.090746	ثلاثه ايام: 0.100586	فيح جهنم: 1.0	سحر: 0.041237
الباس: 1.0	ذي حمه: 0.090746	تربه ارضنا: 0.100586	فوح جهنم: 1.0	رقيه: 0.030927

TABLE 4: The top 15 ranked concept extracted by different algorithms for Medicine domain from Hadith corpus.

3.4 Candidate Concepts Selection using Domain knowledge

After extracting terms based on the different algorithms, there are several important concepts to the domain but have significantly low value. Domain concept extraction can benefit from using domain knowledge, because background concepts with low frequency can be found: even if they occur only once or twice in the given corpus, they will be extracted if they contain domain knowledge units. Our method integrates the domain knowledge to allow extracting the concepts which are semantically relevant to the target domain.

As seen from the Table 4, العود الهندي: 0.333, القسط البحري: 0.333, الحبه السوداء: 0.333, which are important concepts for the domain but they have scores less than the average threshold for the RCF algorithm which is 0.4. After domain knowledge integration, a high score assigned to these concepts and extracted as an important concepts for the domain as shown in Table 5.

3.4.1 New Weight Calculation Based On The Domain Knowledge

After assigning weights to the concept candidates in C, we use domain list to extract the relevant terms. For each of the extracted candidate concepts, we identify the individual words of the phrase. If all of its words or at least one of its words is in the domain list, then the term is assigned high score and defined as a domain concept. When the heuristic {If all of its words are in the domain list} is applied, this will increase precision of the method. While when the heuristic {if at least one of its words are in the domain list}, the recall will be increased.

Give a candidate $c \in C$, its domain relevance measure is defined as follows:

New-Weight(c) = statistical-values (c) + domain-value (c)

Where:
statistical-values (c) is the weight assigned to the concept using the different statistical measures.
domain-value (c) is the weight assigned to the concept using domain list.

Finally, we generate a ranked list of concept candidates {c1,c2,……..cn} ∈ C according to their weights.

Algorithm 3 shows the algorithm for the candidate selection using the domain knowledge. Table 5 shows the top 15 ranked concept extracted using statistical algorithms and domain knowledge.

```
Algorithm: Extract candidate terms and ranking
Purpose: extract candidate terms and ranking using domain knowledge
Input : List of candidate terms Lterms
          and Domain specific list Dlist
Avg-threshold= frequency threshold for candidate terms
Output: Lconcepts: list of ranked concepts
Begin
For each candidate term t in Lterms
     If t exist in Dlist return D-weight(t)
        For each word in t
          If t[word] exists in Dlist  return D-weight(t)
          Weight (t) =S-weight (t)+ D-weight(t)
           Add t to Lconcepts
     End
  Rank Lconcepts by their weight value
  Select from Lconcepts the candidate above AVG-threshold
End
End
```

ALGORITHM 3: The Candidate Selection using The Domain Knowledge.

CF-DF	**TF-IDF**	**CF-IDF**	**RCF**	**Avg-CF**
المن: 2.0	بيدك الشفاء: 1.090746	فيح جهنم: 1.100586	فيح جهنم: 2.0	العذره: 1.030927
العين حق: 2.0	المن: 1.090746	بيدك الشفاء: 1.100586	بيدك الشفاء: 2.0	ذات الجنب: 1.025773
الحمى: 2.0	العين حق: 1.090746	العين حق: 1.100586	المن: 2.0	العود الهندي: 1.025773
فيح جهنم: 1.666666	العسل: 1.090746	المن: 1.085167	العين حق: 2.0	القسط: 1.020618
بيدك الشفاء: 1.25	فيح جهنم: 1.081135	العسل: 1.085167	العسل: 2.0	الشفاء: 1.015463
المبطون: 1.166666	الحمى: 1.074317	الحمى: 1.069748	الحمى: 2.0	الحمى: 1.015463917
الشفاء: 1.15	للمريض: 1.045373	للمريض: 1.042583	للمريض: 1.5	الحبه السوداء: 1.015463
العذره: 1.010309	المبطون: 1.045373	المبطون: 1.042583	المبطون: 1.5	الطاعون: 1.012886
العود الهندي: 1.008591	الكحل: 1.045373	الكحل: 1.042583	الكحل: 1.5	فيح جهنم: 1.010309
الطاعون: 1.007363	الدم: 1.0453731	الدم: 1.0425836164723705	العين: 1.5	العين: 1.010309
القسط: 1.006872	العين: 1.040567	البان الاتن: 1.040234	الشفاء: 1.5	السحر: 1.010309
ذات الجنب: 1.006443	الشفاء: 1.037158	العين: 1.038073	الدم: 1.5	الحبيبه السوداء: 1.010309
العين: 1.005154	البان الاتن: 1.036298	الشفاء: 1.034874	البان الاتن: 1.4	وجهه الدم: 1.005154
العسل: 1.005154	القسط البحري:	القسط البحري: 1.033528	القسط البحري:	للمريض: 1.005154

	1.0302487		1.333333	
الحبه السوداء: 1.005154	القسط: 1.027045	العود الهندي: 1.033528	القسط: 1.333333	بيدك الشفاء: 1.005154

TABLE 5: The top 15 ranked concept extracted using domain knowledge for Medicine domain from Hadith corpus.

We observed that our method extracted multi-word terms as concept candidates, while the others tend to extract single word terms. For example, in Table 4 looking at the 15 concept candidates extracted by Avg-CF algorithm, most terms are single word terms. On the other hand, in Table 5 we observed that our method using domain specific knowledge able to extract multi-word terms that are relevant concepts.

3.5 Post Processing

Refinement and Validation

An important task of this stage is the removing of unrelated concepts from the extracted candidate concepts list. We adopt a pruning strategy: terms that are frequently used in a corpus are likely to denote domain concepts, while less frequent terms removed from the extracted candidate concepts list.

We set an average frequency of the extracted concept as a threshold value and prune all concepts that have a frequency lesser than this value.

$$average\ concept\ frequency = \frac{\sum_{i=1}^{n} f(ci)}{n} \qquad (10)$$

Where n is the total number of concepts.

When applying threshold average, the remaining concepts omitted from candidate list which reduce recall but increase precision, if we need to keep all terms we don’t check for threshold. As not all terms generated by our methods are domain concept, the domain experts determine their relevance to the domain. The relevant concepts are then used to represent domain concepts.

4. EXPERIMENTS, EVALUATION AND RESULTS

4.1 Experiments

We setup a number of experiments to investigate the effectiveness of our method using the different statistical measures and domain knowledge for Arabic concept extraction. In the first experiment, we implemented and compared the following algorithms.

TF-IDF(as baseline) and modified version from it for multi-word terms CF-IDF, RCF, Avg-CF and our domain relevance measure CF-DF to rank the candidate concepts. In the second experiment, we integrated domain knowledge to assign weights for the candidate concept

In this section, the experimental results obtained by our method are presented. As mentioned above, the experiment has been conducted in the Medicine Book from Hadith corpus. In order to evaluate the performance of our method, recall and precision scores will be calculated in these experiments. These measures are the most commonly used for the assessment of terms extraction systems, and trace their origins back to the Information Retrieval discipline.

Precision is defined as the number of concepts correctly returned by the extractor, divided by the total number of concepts returned by the extractor (see equation 11). On the other hand, recall is the ratio of the number of concepts correctly extracted to the total number of concepts in the documents (see equation 12).

$$Precision \;= \frac{\text{correct concepts extracted}}{candidate\ concepts\ extracted} \qquad (11\ )$$

$$Recall \;= \frac{\text{correct concepts extracted}}{concepts\ in\ the\ documents} \qquad (12\ )$$

To find the number of concept correctly identified by our extractor, domain experts are needed to examine the output. For recall, to find the total number of concepts in the documents, domain experts are needed to identify them manually from documents. Because manually identifying concepts from documents is time consuming, in these experiments we followed the previous studies in [1], we used precision only as a measure to evaluate our method.

4.2 Results

For each of the five experiments, we list the top 15 ranked concepts from medicine field from hadith corpus. The ranked candidate concepts are extracted by the measures CF-DF, TF-IDF, CF-IDF, RCF and Avg-CF.

For evaluating our method, we tested the precision of the five measures and the algorithms after domain knowledge integration. The results are shown in Tables 6 and 7 respectively.

Since evaluating entire candidate lists is tedious, past studies have focused on the top-n terms [1]. Previous research has also shown that the evaluation over a sample of size n are comparable to the evaluation over the entire population set [15]. So, we have calculated the precision for each of the above experiments. For each experiment, the concepts to be extracted set to top 100, top 200 and above average threshold.

The precision values of the candidates extracted at the different values N were computed using equation 11.The comparisons among the performances of each algorithm have been shown in Table 6 . And the performances after domain knowledge integration have been shown in Table 7.

Total number of concepts evaluated = N					
N	**CF-DF**	**TF-IDF**	**CF-IDF**	**RCF**	**Avg-CF**
>average threshold	0.92	0.65	0.65	0.67	0.66
Top 100	0.88	0.81	0.81	0.72	0.69
Top 200	0.72	0.66	0.65	0.68	0.60

TABLE 6: The precision for each algorithm evaluated for Medicine domain from Hadith corpus.

The reported results are as follows: for CF-DF, TF-IDF, CF-IDF, RCF, Avg-CF the precision was 0.92, 0.65, 0.65, 0.67, 0.66 respectively at N above threshold value. As shown in the Table 6, the method using the CF-DF performs the best for all the Ns. The TF-IDF and CF-IDF measures produced the same precision at all N. This is due to that the important concepts in the domain are not appearing frequently in most of the domain documents. The other RCF, Avg-CF measures show fluctuations of performance at different N.

Table 7 shows the evaluation results for all algorithms after domain knowledge integration. We can observed the increase in the precision for all algorithms at different N. for CF-DF, TF-IDF, CF-IDF, RCF, Avg-CF the precision was 0.94, 0.88, 0.90, 0.74, 0.97 respectively at N above threshold value. From the experiment, we can conclude that, the CF-DF algorithm produced the best and most stable results before and after domain knowledge integration. Important concepts concentrate on the top of the list, while non important concepts tend to appear at the bottom of the list. The Avg-CF algorithm is the best in the concept coverage.

Most of the studies described in the previous section for multiword terms extraction from Arabic text 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. Compared to other method our method consider contextual information using domain knowledge which increase precision compared to 85% for [5], 80% for [26] and 82% for [27].

Total number of concepts evaluated = N					
N	**CF-DF**	**TF-IDF**	**CF-IDF**	**RCF**	**Avg-CF**
>average threshold	0.94	0.88	0.90	0.74	0.97
Top 100	0.90	0.88	0.85	0.82	0.75
Top 200	0.77	0.74	0.73	0.75	0.65

TABLE 7: The precision for each algorithm evaluated after using domain knowledge.

Figure 2 illustrates the comparison between the five algorithms for concept extraction, and Figure 3 illustrates the comparison between the five algorithms after domain knowledge integration.

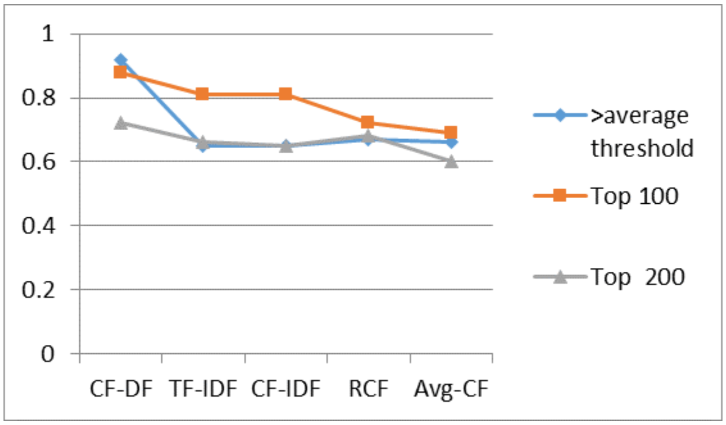

FIGURE 2: The performance of algorithms for concept extraction at different N.

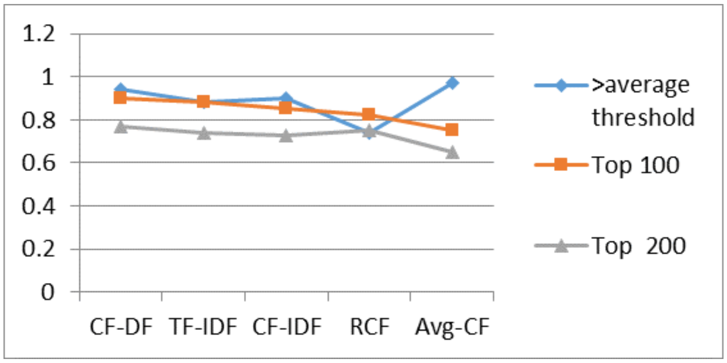

FIGURE 3: The performance of algorithms after using domain knowledge at different N.

As can be shown from experimental results, CF-DF outperforms all methods in term of precision.
To see how our method performs in different domains, we compared the performance of the method on medicine domain and food domain. Table 8 shows its precision in both Medicine domain and food domain when the number of extracted concepts is 100. The results show that our method for concept extraction performs better in Medicine domain than in food domain. Some errors were created due to errors of the subsequent POS tagging and noun phrase extraction. In Medicine field the noun phrases is more than in food field. While food domain has more verbs and the tagger tags it as noun. Accordingly, the method will assign a high weight to the terms and considered as domain concept. For example, 1.0 :الدباء ياكلها , 1.0 :واكل الاقط.

The results (Table 8) show that the precisions in the range of 55% and 65%. This means that more than half of the concepts are identified. Most of the concepts are single word terms and it is identified correctly by our method. For example, 1.02 :الدباء, 1.0 :التمر, 1.0 :الثريد, 1.0 :الثوم. For multi word concept, 0.088 :خبز بر مادوم , 0.088 :انيه الذهب والفضه, 1.5 :دباء وقديد.

Total number of concepts evaluated = 100					
	CF-DF	TF-IDF	CF-IDF	RCF	Avg-CF
Medicine	0.90	0.88	0.85	0.82	0.75
Food	0.60	0.55	0.57	0.60	0.63

TABLE 8: The precision for each algorithm evaluated for medicine domain and Food domain from hadith corpus.

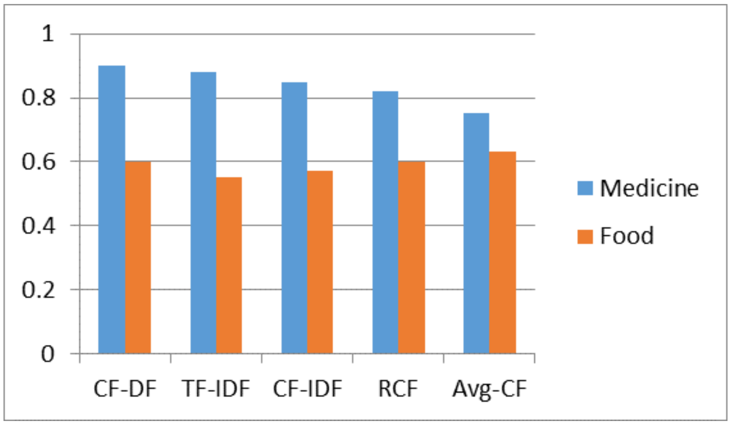

FIGURE 4: The performance of algorithms for concepts extraction in Medicine domain and Food domain.

5. CONCLUSION

We have presented a new method for concept extraction from Arabic texts. It uses a combination of linguistic, statistical and domain knowledge to extract domain relevant concepts. We proposed what NLP techniques are used to extract concept candidates, and how statistical and domain knowledge can be used and combined to extract domain relevant concepts.

Our method performs in unsupervised manner which means, no need for training data. Also, it does not require general corpora to measure relevance of concepts to the domain. This leads to avoiding the problem of skewness in terms frequency information. Further, we experimented with a small set of documents and we have the promising results, this is different from other methods that is work effectively only for large number of documents in the corpus.

Our method also used effectively for concept extraction in various domains. Although it works by combining statistical and domain knowledge, but our evaluation showed that statistical information can be used alone if no domain knowledge available.

The various experiments have been performed to assess the effectiveness of each used algorithm. We reported in the previous section the experimental results on concept extraction from Arabic texts. The experimental results show that the concept extractor module is effective in extraction concept from Arabic documents. Our method after domain knowledge integration performed better than the algorithms used alone.

In conclusion, our initial experiments support our assumption about the usefulness of our method for concept extraction. As shown through the evaluation, our method has a strong ability to extract domain relevance concept using a combination of statistical measures and domain knowledge. This overcomes the problems that have been found in the methods that based only on the frequency or TF-IDF to measure the importance of the candidates. From our initial results, we have found that using domain knowledge to determine domain relevant concept increases the precision of concept extraction.

Our contributions to the Arabic concept extraction field are as follows. We evaluate and compare different statistical measures and proposed a new one for candidates weighting. Our method can extract rare concepts, even those appearing with low frequency. It also excludes irrelevant concepts even if they occur frequently in the corpus.

To see how our method performs in different domains, we compared the performance of the method on other domains. In the food field results, the precision seems to be worse than that in the medicine field. This result is because of the errors of the subsequent POS tagging and noun phrase extraction. On the other hand, more than half of the concepts are correctly identified.

The results show the high effectiveness of the proposed approach to extract concepts for Arabic ontology construction. The output list from this module constitutes the fundamental layer of ontologies. In the future, we will continue to evaluate and compare results of other domains and we will use other preprocessing tools to enhance the precision. We will develop a method for semantic relation extraction between the extracted concepts.

6. REFERENCES

[1] A.Al-Arfaj and A. Al-Salman. “Towards Concept Extraction for Ontologies on Arabic language,” in Proceeding of 3rdInternationalConference on Islamic Applications in Computer Science And Technology, 1-3 Oct 2015, Turkey.

[2] A.Al-Arfaj and A. Al-Salma. “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, pp. 246 – 251.

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

[4] A. Zouaq, D. Gasevic, M. Hatala. “Towards open ontology learning and filtering”. Information Systems, vol. 36, no.7, pp. 1064–1081, 2011

[5] S. Boulaknadel, B. Daille and D. Aboutajdine. "A multi-word term extraction program for Arabic language," in Proceeding of the 6th International Conference on Language Resources and Evaluation, May 28-30, Marrakech Morocco., 2008, pp.1485-1488.

[6] I.Bounhas and Y.Slimani. "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, August 21-23 2009, pp. 429-436.

[7] P. Buitelaar, P. Cimiano and B. Magnini. “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 2005.

[8] P. Cimiano and J. Volker. "Text2Onto – A Framework for Ontology Learning and Data Driven Change Discovery," in Proceeding of the 10th International Conference on Applications of Natural language to Information system (NLDB), Spain, 2005, pp. 227–238.

[9] P. Cimiano., A. Madche., S. Staab and J. Volker. “Ontology Learning.” IN Staab,S and Studer, R (eds.), Handbook on Ontologies, International Handbooks on Information

Systems, DOI 10.1007/978-3-540-92673-3, Springer-Verlag Berlin Heidelberg 2009, pp.245-267.

[10] B. Daille. “Study and Implementation of Combined Techniques for Automatic Extraction of Terminology.”. In Resnik and Judith (Eds): The Balancing Act: Combining Symbolic and Statistical Approaches to Language, Cambridge, MA, USA: Mit Press, 1996, pp.49-66.

[11] S. El-Beltagy and A. Rafea. "KP-Miner: A Keyphrase Extraction System for English and Arabic Documents". Information systems, 34(1), pp. 132-144, 2009.

[12] X. Jiang and A. Tan. "CRCTOL: A Semantic-Based Domain Ontology Learning System." Journal of the American Society for Information Science and Technology (JASIST), 61(1), pp.150-168, 2010.

[13] R. Navigli and P. Velardi. “Learning Domain Ontologies from Document Warehouses and Dedicated Websites,” Computational Linguistics, 30(2), pp. 151-179, 2004.

[14] N. Noy and D. McGuinness. "Ontology Development 101: A Guide to Creating Your First Ontology." Report SMI-2001-0880, Department of Mechanical and Industrial Engineering, University of Toronto,pp.1-25, 2001

[15] M. Pazienza., M. Pennacchiotti and F. Zanzotto. “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, 2005, pp. 255–279.

[16] J. Qiu., Y. Chai., Y. Liu., Z. Gu., S. Li and Z. Tian. “Automatic nontaxonomic relation extraction from big data in smart city,” IEEE Access, vol. 6, pp. 74854–74864, 2018

[17] J. Qiu., Y. Chai., Z. Tian., X. Du and M. Guizani. “Automatic Concept Extraction Based on Semantic Graphs From Big Data in Smart City,” IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, pp.1-9,2019.

[18] A.Saif and M. AbAziz, "An Automatic Collocation Extraction from Arabic Corpus." Journal of Computer Science ,7 (1), pp. 6-11, 2011.

[19] K. Toutanova,, D. Klein., C. Manning and Y. Singer. “Feature-rich part-of-speech tagging with cyclic dependency network,” In Proceedings of Human Language Technology –North American Chapter(HLT-NAACL), 2003, pp. 252–259.

[20] S. Tulkens., S. Suster and W. Dealemans. “Unsupervised concept extraction from clinical text through semantic composition”. Journal of Biomedical Informatics. Vol.91,pp.103-120,2019

[21] Z. Zhang., B. Christopher and F. Ciravegna. “A Comparative Evaluation of Term Recognition Algorithms,” in Proceeding of The 6th International Conference on Lnaguage Resources and Evaluation (LREC2008), May 28-31,2008, Marrakech, Morocco, pp.2108- 2113.

[22] S. Zaidi., M. Laskri and A. Abdelali. "Arabic collocations extraction using Gate," in Proceeding of International Conference on Machine and Web Intelligence (ICMWI), 2010, pp. 473 - 475.

[23] M. Rizoiu and J. Velcin. “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.).2011, pp. 38-61

[24] Z. Harris. “Distributional structure”. structural and transformational linguistics, pp.775–794, 1970.

[25] K. Sarkar. “A Hybrid Approach to Extract Keyphrases from Medical Documents.” International Journal of Computer Application, 36(18), pp. 14-19, 2013.

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

[27] A. El-Mahdaouy, S. Alaoui Ouatik and E. “ A study of association measures and their combination for Arabic MWT extraction” in Proceedings 10th International Conference on Terminology and Artificial Intelligence,2013, pp. 45-52.