An Efficient Semantic Relation Extraction Method For Arabic Texts Based On Similarity Measures

Abeer Alarfaj *aaalarfaj@pnu.edu.sa*
Department of Computer Science
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

Semantic relation extraction is an important component of ontologies that can support many applications e.g. text mining, question answering, and information extraction. However, extracting semantic relations between concepts is not trivial and one of the main challenges in Natural Language Processing (NLP) Field. In this paper, we propose a method for semantic relation extraction between concepts. The method relies on the definition of concept context and the semantic similarity measures to extract relations from domain corpus. In this work, we implemented algorithm for concept context construction and for similarity computation based on different semantic similarity measures. We analyze the proposed methods and evaluate their performance. The preliminary experiments showed that the best results precision of 83% are obtained with Lin measure at minimum confidence =0.50 and precision of 85% with the Cosine and Jaccard similarity measures. The main advantage is the automatic and unsupervised operation; it doesn't need any pre labeled training data. Also used effectively for relation extraction in various domains. The results show the high effectiveness of the proposed approach to extract relations for Arabic ontology construction.

Keywords: Relation Extraction, Arabic NLP, Arabic Semantic Relation Extraction, Concept Context, Semantic Similarity Measures.

1. INTRODUCTION

Relation extraction is an important aspect of ontology construction. Relation learning defined by Cimiano as "a task of learning relation identifiers or labels r as well as their appropriate domain and range" [1]:

In the current research concerning non-taxonomic relation extraction, the existing approaches can be classified into the following:

- Statistical approach relies on the distributional properties of words through co-occurrence distribution of words. In order to extract the correlated concept pairs the semantic distances between words are computed.
- Lexico-syntactic approach relies on patterns matching based on syntactic structure to extract non-taxonomic relations between concepts.
- Hybrid Approach

However, linguistic-based techniques using static rules tend to face difficulties in coping with the structural diversity of a language. In order to identify indirect relations, statistics-based techniques such as co-occurrence analysis are necessary. The current approaches use both pattern matching and statistical analysis based on co-occurrence.

The Arabic language compared with the English language has a much more complex syntax. So, the need for new methods to construct ontology from Arabic texts is growing. The Arabic ontology is a necessary knowledge for applications that process Arabic documents [2].

In order to extract relations between concepts, different techniques from machine learning and natural language processing community have been applied in ontology learning.

The main contributions of this work are the following:

1. The new semantic relation extraction methods.
2. The new context definition of concepts based on relevance analysis.
3. Similarity measures (Cosine, Jaccard, Dice, and Lin) between concepts vectors.
4. Construction of an initial taxonomy using a seed concepts and noun-phrase based patterns.

In this paper, each contribution to semantic relation extraction is described and illustrated with examples. After the introduction, section 2 present semantic relation extraction approaches from Arabic texts. In section 3, we introduce our approach to semantic relation extraction. In section 3.1, the key details of the proposed method for relation extraction is presented. In section 3.2, we discuss the algorithm of concept context extraction. In section 3.3, the proposed method for semantic similarity computation and the measures of semantic similarity are discussed. In section 4, the automatic ontology construction method is presented. In Section 5, the experimental results are presented and discussed. Finally, Section 6 concludes the paper.

2. SEMANTIC RELATION EXTRACTION

A popular approach to relation extraction from Arabic text is based on the lexico-syntactic patterns. The authors in [3] used an enhanced version of Hearst's pattern to an Arabic corpus. Their enhanced algorithm include: pattern enrichment, pattern filtering, the application of negative patterns and pattern evaluation. Their evaluation results reached 78.57% average precision and 80.71% average recall. [4] presented a pattern-based and seed ontology method for extraction of antonyms from Arabic corpus. The extracted patterns then used to discover new antonym pairs to enrich ontology. [5] proposed a semi supervised pattern based bootstrapping technique to extract semantic relations between entities. They experimented their method with two corpora which differ in size and genre, reaching a highest F measure of 75.06%. The main drawbacks of these approaches are complexity of pattern construction and the low recall. Also, the implicit relations are missing and only explicit relations are extracted by these patterns. Another studies for Arabic used the statistical approach that based on co-occurrence technique and machine learning algorithms to detect and classify the relations [7,8,9], do not require any manual labor, but it tends to generate a large number of relations. [6] proposed a relation extraction algorithm based on MaxEnt classifier, which resulted in 85% accuracy. Hybrid approaches combine statistical learning with linguistic knowledge and takes the advantages of both [10]. A distributional approach for calculating similarities proposed in [11], which is based on syntactic dependencies to extract semantic relations. They achieved 60% as the most decreased rate compared to 67% as the best result for the co-occurrence method.

2.1 Taxonomy Extraction

The taxonomy extraction approaches are based on lexico-syntactic patterns or hierarchal clustering methods. The main drawback of the approaches based on lexico-syntactic patterns is low recall and for the clustering based approaches it is difficult to label clusters. So, in this work we address the problems of the existing approaches and we propose a method for extracting taxonomic relations between concepts using terms compound structure. The taxonomic relations are extracted from the compositions of terms. Using the composition of the concepts, we can extract hierarchal relations between concepts. For compound terms, head words are extracted and taxonomic relation between the head word and the compound terms is extracted. For example, for concepts القسط الهندي is subclass from القسط , and more recursive القسط الهندي والبحري is subclass from القسط الهندي. الحجامه من الداء is sub class from الحجامه . Taxonomic relations extracted using compound term heuristics are shown in Figure 1.

When we compare them to the Hearst's patterns, with this approach we only are able to obtain a reduced subset of the possible hyponyms for a domain but its simplicity results in a higher robustness.

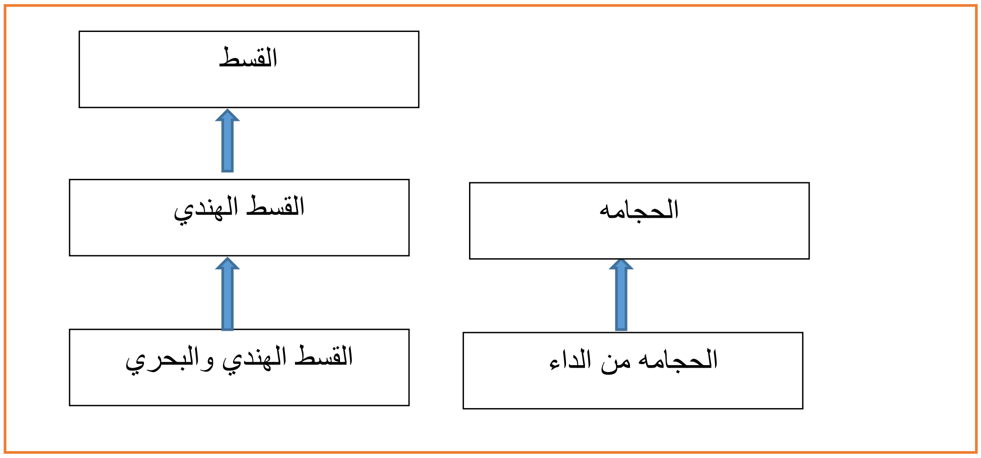

FIGURE 1: Compound Term Method For Taxonomy Extraction.

3. THE PROPOSED METHOD FOR RELATION EXTRACTION

Our proposed method is twofold. In a first step, we extract context for seed and candidate concepts. In a second step, we extract semantic relations between the extracted concepts and seed concepts using similarity measures. In this section, we describe our method for context extraction and relation extraction before presenting our method for automatic ontology construction. A detailed description of the developed semantic relation extraction methods is presented. The proposed method for relation extraction is shown in Figure 2.

The steps of our approach are as follow:

The input is seed list and concept list and domain documents after filtering to reduce computation.

1. Extract context for each seed concept
2. Extract context for each candidate concept
3. Create vector for each seed concept and candidate concept
4. Compute semantic Similarity between each candidate concepts and seed concepts
5. Assign candidate concepts to the seed concept with the highest similarity.

In this approach, the algorithm to extract contextual information associated with concepts is implemented. The algorithm that compares and measures context vectors exploiting semantic similarity between concepts and candidate concepts is also implemented.

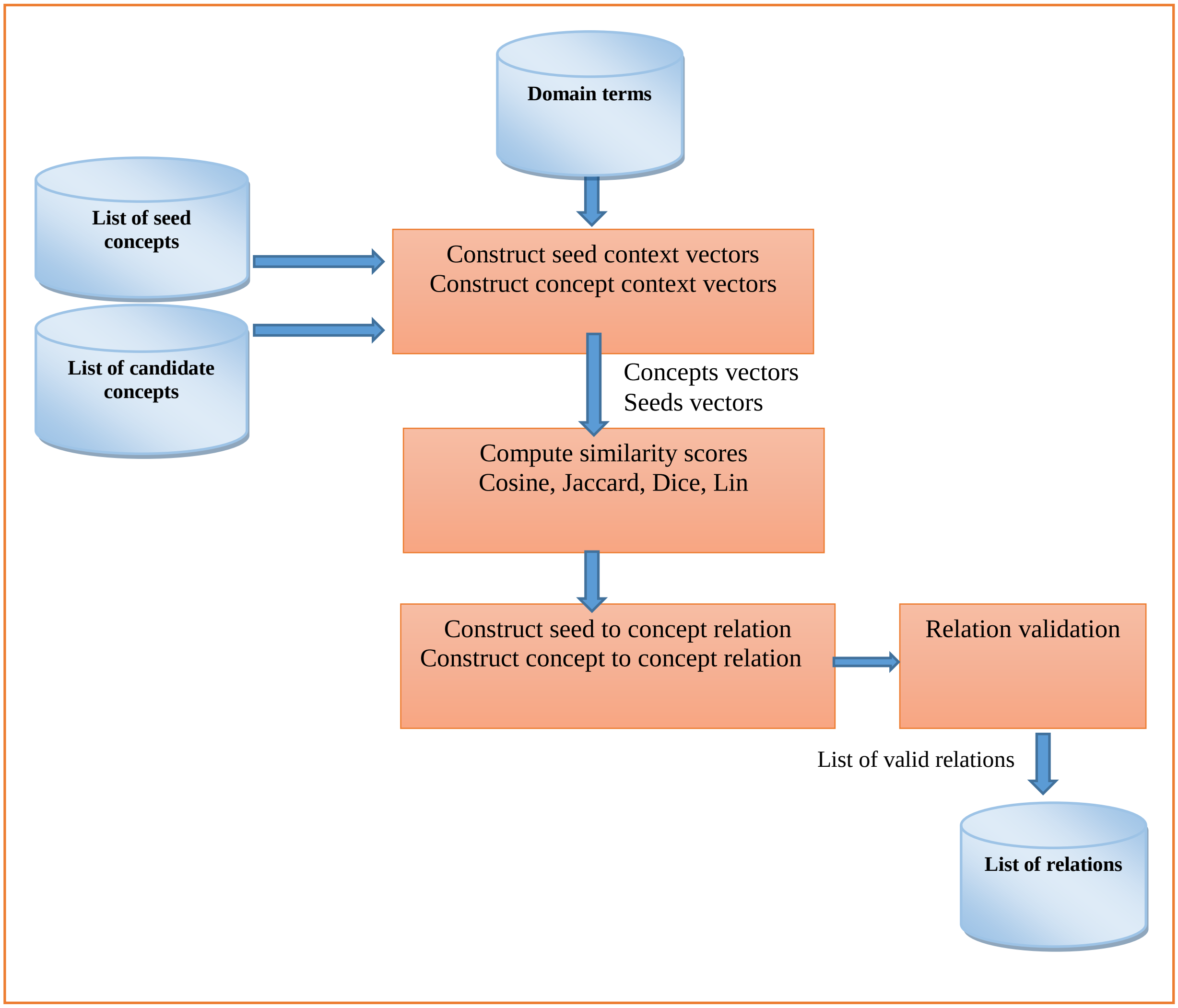

FIGURE 2: The Proposed Relation Extraction Method.

3.1 Algorithm To Construct Concept Context Vector

The idea is to create for each concept a vector with the terms that are strongly connected with it using confidence. So, even if the frequency of term is low and has a strong relation with the concept it will be appear as candidate concept. The main difference though is in defining the notion of context of two concepts, and the measures of their similarities. Based on the distributional hypothesis, "semantically similar words occur in similar contexts". The context construction based on the hypothesis “semantically similar concepts have similar environment”. We formalize this intuition by defining concept environment as all concepts related to it using confidence measure.

Our algorithm for concept context construction is based on Equation (1) which measures the degree of co-occurrence between candidate concept (candidate) and each seed concepts in the domain (seed). The confidence is calculated by the equation:

$$confidence(seed, candidateconcept) = \frac{(seed \cap candidate)}{candidate} \qquad (1)$$

Where:
$(\text{seed} \cap \text{candidate})$ denotes the number of co-occurrence of seed and candidate concept, candidate denotes the number of candidate concept alone. With these computed measures, a context vector for all concepts and seed is constructed. The threshold (min-conf) is applied for discarding low degree of co-occurrence before building context vectors. Table 1 shows the application of the confidence measure on candidate concepts to measure the strength of association.

Let us look closer at the definition above by taking an example. From Table 2, the vector for concept (الشفاء) can be constructed as:

Vec(الشفاء)={الكي ,1.0, امتي ,1.0, شربه عسل ,0.5, كيه نار ,1.0, شرطه محجم ,0.4}.

Algorithm 1 shows the concept context construction algorithm.

The example	**confidence measure**
Candidate count(فيح جهنم)=3.0	confidence = 0.6
Candidate-seed count(الحمى, فيح جهنم)= 2.0	
Candidate count(القسط البحري)=2.0	confidence = 0.5
Candidate-seed count(القسط, القسط البحري)= 1.0	
Candidate count(القسط)=3.0	confidence = 0.0
Candidate-seed count(الطاعون, القسط)= 0.0	

TABLE 1: The Example of Confidence Measure.

Candidate Concept (الشفاء)	**Concept**	**Confidence**
	الكي	1.0
	امتي	1.0
	شربه عسل	0.5
	كيه نار	1.0
	شرطه محجم	0.4

TABLE 2: The Vector For Concept (الشفاء).

```
Algorithm for concept and seed context vector building
Input: list of seed concepts, list of candidate concepts
Output: Seed(s1,{cand1,conf,cand2,conf,cand3,conf})
          Concept(c1,{cand1,conf,cand2,conf,cand3,conf})
For each s in S seed concepts {
   For each c in C candidate concepts {
      If s != c concept
        For all documents in the corpus {
          Compute occurrence seed-cand-count(seed and candidate)
          Compute occurrence can-count(candidate)
        }
Compute confidence= seed-cand-count/ cand-count
 If confidence >threshold
     put in seed-context(s,{c, conf})
    }
  }
}
```

ALGORITHM 1: Context Construction.

3.2 The Proposed Method For Semantic Similarity

Algorithms of semantic relation extraction proposed in this module, are based on the hypothesis, which states that semantically similar concepts have similar environment. [12] defines the environment of a concept to be the set of concepts related to it. Figure 3 shows the equivalent concepts common between two concepts. The method takes as an input a set of concepts and seed concepts and outputs a set of relations between them. For finding the related concepts to the candidate concepts, we compute semantic similarity between concepts and seed by computing similarity measure for their context vector.

For each seed concepts, we computed an average value AVG *(similarity -score)* by equation (2) for all candidate concepts. We used AVG *(similarity*-score*)* as the threshold value for candidate concepts. We compute average similarity for seed concept using the equation:

$$Avg(seed\ i) = \frac{1}{n}\sum_{j=1}^{n} sim(seed\ i, candidate\ j) \qquad (2)$$

Where n is the number of candidate concepts related to seed.

These candidates then sorted according to the decreasing values of their similarity scores and keep only the candidate concepts above the AVG *(similarity*-score*)*. The algorithm is shown in Algorithm 2.

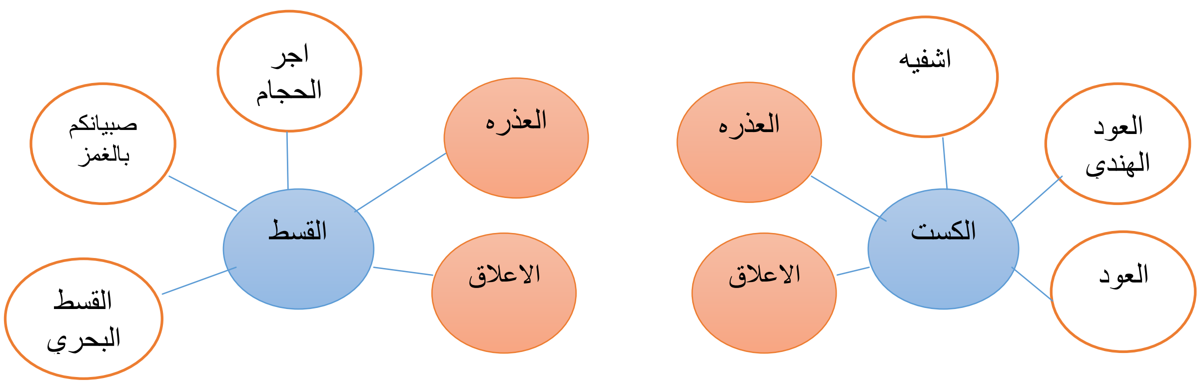

FIGURE 3: Concepts that co-occur with concept القسط and الكست

```
Algorithm for similarity computation between seeds and
concepts
Input: seeds vectors, concepts vectors
Output: seed-concept similarity scores
For each s in seed vector list {
    For each c in concept vector list {
        If seed!= concept {
         Compute similarity between seed and concept vectors
         similarity-score(s,c) = Sim(Vec(s),Vec(c)
         similarConcept.put(c, similarity-score);
    }
  }
         Comput average threshold for each seed similarity-scores
         If similarity-score >average threshold
          Avg-similarConcept.put(c, similarityscore)
      Sort (Avg-similarConcept)
      seedconceptRel.put( s,{c, similarityscore})
}
```

ALGORITHM 2: Similarity Computation between Seeds and Concepts.

3.2.1 Similarity Measures

Different similarity measures and their evaluation are available from the statistical natural language processing community. The measures within our work, namely the Cosine similarity, Jaccard index, Dice Coefficient and Lin measures are briefly introduced. We selected them because each of the measures was well respected in lexical semantic similarity field [13,14,15].

Cosine Similarity Measure

The cosine measure between two vectors measures the similarity in terms of comparing the cosine angle between two vectors. The less the angle, the higher the similarity. Formally, if $\vec{A}$ and $\vec{B}$ are two concept vectors, then their similarity is computed as:

$$Sim(A,B) = Cos(A,B) = \frac{A.B}{|A||B|} \tag{3}$$

$$\mathrm{Cos}(A,B) = \frac{\sum_{i=1}^{n} AiBi}{\sqrt{\sum_{i=1}^{n} A_i^2}\,\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{4}$$

Where: $|A|$ *and* $|B|$ *are the lenghts of A and B vectors.*

A_i and B_i are components of vector A and B respectively.

The resulting similarity ranges from −1 meaning exactly opposite, to 1 meaning exactly the same, with 0 indicating orthogonality (decorrelation), and in-between values indicating intermediate similarity or dissimilarity.

Jaccard Index

The Jaccard Index, also known as the Jaccard similarity coefficient, measures similarity in terms of the relation between the intersection and union of two sample sets:

Let A and B be two sets Jaccard coefficient:

$$\mathrm{Jaccard}(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{5}$$

$0 \leq Jaccard(A,B) \leq 1.$

If A and B are two vectors then their Jaccard similarity coefficient is defined as:

$$\mathrm{Jaccard}(A,B) = \frac{\sum_i \min(A_i, B_i)}{\sum_i A_i + \sum_i B_i - \sum_i \min(A_i, B_i)} \tag{6}$$

Dice's Coefficient

Dice's Coefficient has a lot in common with the Jaccard Index, but weights matching's twice, compared to The Jaccard Index. Dice's Coefficient measures similarity over sets in the given way:

$$\mathrm{Dice}(A,B) = \frac{2|A \cap B|}{|A| + |B|} \tag{7}$$

For the similarity between vectors we can use:

$$\mathrm{Dice}(A,B) = \frac{2\sum_i \min(A_i, B_i)}{\sum_i A_i + \sum_i B_i} \tag{8}$$

Lin's Measure

Lin's similarity measure is based on the information content of each concept. The more common information two concepts share, the more similar the concepts are. It uses both the amount of information needed to state the commonality between the two concepts and the information needed to fully describe these terms [17].

The measure can be defined as:

$$\text{Lin (A,B)}= \frac{\sum_i (A_i + B_i)}{\sum_i A_i + \sum_i B_i} \quad (10)$$

In this work, we experiment with different similarity measures Cosine similarity, Jaccard index, Dice Coefficient and Lin measure. We investigated the relation extraction method with the semantic similarity measures described above and with various numbers of confidence threshold. The number of extracted relations depends little on the similarity measure type. Table 3 shows different similarity measures scores for concepts with minimum confidence =0.25 and Table 4 shows different similarity measures scores for concepts from Medicine domain with minimum confidence =0.50. Tables 5 and 6 show different similarity measures scores for concepts from Food domain with minimum confidence =0.25 and 0=.50 respectively. Table 7 shows different similarity measures scores for concepts from Good Manners domain with minimum confidence 0=.50.

All concepts related to = الحبه السوداء	Cosine similarity score	Jaccard similarity score	Dice similarity score	Lin similarity score
داء	0.93478	0.862745	0.92634	0.93032
السام	0.91499	0.843137	0.91924	0.91924
الحبه	0.91499	0.843137	0.91924	0.91924
زيت	0.82189	0.590604	0.76983	0.81304
انفه بقطرات	0.82189	0.590604	0.76983	0.81304
الموت	0.82189	0.590604	0.76983	0.81304
الحبيبه السوداء	0.82189	0.590604	0.76983	0.81304
الحبيبه	0.82189	0.590604	0.76983	0.81304
شفاء	0.64527	-	0.57471	0.58065
الجانب	0.82189	0.590604	0.76983	0.813040

TABLE 3: Similarity measures scores for الحبه السوداء concept with min-conf =0.25 from Medicine domain.

All concepts related to = الحبه السوداء	Cosine similarity score	Jaccard similarity score	Dice similarity score	Lin similarity score
داء	0.92347	0.92347	0.92631	0.92631
السام	0.90567	0.90567	0.91489	0.91489
الحبه	0.90567	0.90567	0.91489	0.91489
زيت	0.81723	0.81723	0.74261	0.77637
انفه بقطرات	0.81723	0.81723	0.74261	0.77637
الموت	0.81723	0.81723	0.74261	0.77637
الحبيبه السوداء	0.81723	0.81723	0.74261	0.77637
الحبيبه	0.82189	0.81723	0.74261	0.77637
شفاء	0.81723	-	0.59119	-
الجانب	0.81723	0.81723	0.74261	0.776371

TABLE 4: Similarity measures scores for الحبه السوداء concept with min-conf =0.50 from Medicine domain.

All concepts related to **الدباء**	Cosine similarity score	Jaccard similarity score	Dice similarity score	Lin similarity score
دباء وقديد	0.64180	0.36719	0.53715	0.59534
القصعه	0.54226	0.25091	0.40117	0.45009
خبز	0.36798	0.17877	0.30331	0.36966
الثريد	0.32291	0.18222	0.30827	0.33834
الطعام	0.16988	0.09065	0.16624	0.20654
شعير	0.16404	0.08730	0.16058	0.19951

TABLE 5: Similarity measures scores for **الدباء** concept with min-conf =0.25 from Food domain.

All concepts related to الدباء	Cosine similarity score	Jaccard similarity score	Dice similarity score	Lin similarity score
دباء وقديد	0.56511	0.29729	0.45833	0.47916
القصعه	0.48786	0.21621	0.35555	0.37777
خبز	-	-	-	-
الثريد	0.40451	0.16216	0.27906	0.27906
الطعام	-	-	-	-
شعير	-	-	-	-

TABLE 6: Similarity measures scores for الدباء concept with min-conf =0.50 from Food domain.

All concepts related to الكبائر	Cosine similarity score	Jaccard similarity score	Dice similarity score	Lin similarity score
قول الزور	0.70710	0.5	0.6666	0.66666
شهاده الزور	0.70710	0.5	0.6666	0.66666
الشرك بالله	0.70710	0.5	0.6666	0.66666
الرجل والديه	0.54772	-	0.46153	0.46153
الرجل ابا الرجل	0.54772	-	0.46153	0.46153

TABLE 7: Similarity measures scores for concept الكبائر with min-conf =0.50 from Good Manners domain.

4. ONTOLOGY CONSTRUCTION

The automatic ontology construction algorithm starts from the seed concepts by constructing context vector for each seed concepts. Then a context vector for each candidate concept is built. For each candidate concept, we compute similarity between candidate concept vector and seed concept vector. The concepts related to the seed are considered as seed and find all related concepts to it. The output will constitute the ontology (see Algorithm 3). Figure 4 describes the process for ontology construction for seed **صفر**. As proposed in our framework described in [2], the expert checks the final ontology and makes the necessary corrections in order to detect implicit relationships, avoid redundancies. Figure 5 illustrates the part of the ontology obtained from the Medicine documents.

```
Algorithm for ontology construction
Input: list of seed concepts, list of candidate concepts
Output: Related concepts of the domain

Build seed context vector
Build concept context vector

Compute similarity between seed and concept vectors
Compute similarity between concept and concept vectors
For each s in seed-concept-rel {
  For each c in seed-concept-rel {
   Assign to c all related c in concept-concept-rel
}
}
```

ALGORITHM 3: Ontology Construction.

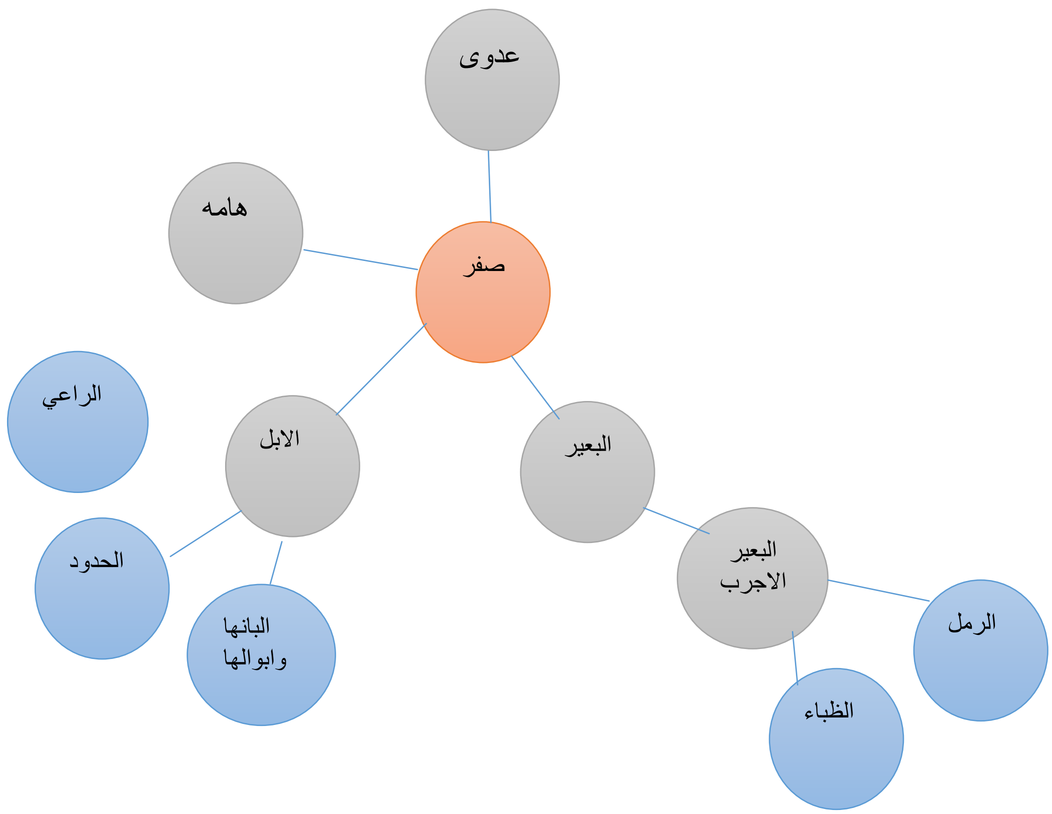

FIGURE 4: Ontology Construction Process for seed **صفر**

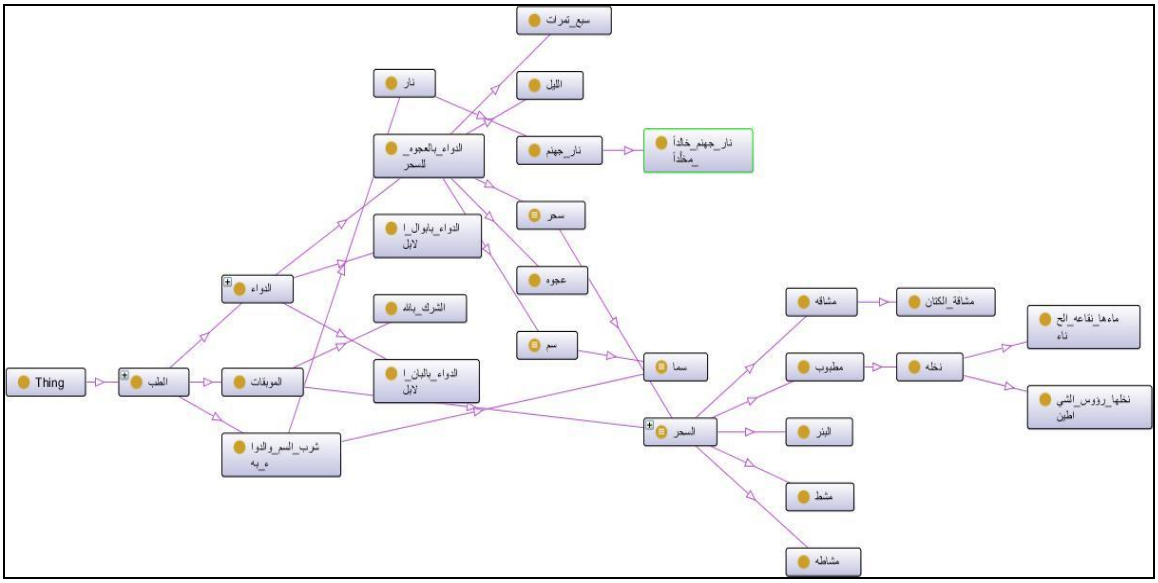

FIGURE 5: Part from The Constructed Ontology for Medicine Domain.

5. EXPERIMENTS, EVALUATION AND RESULTS

5.1 Experiments

We investigated the relation extraction algorithm with the semantic similarity measures described above and with various numbers of confidence threshold (see Tables 3,4,5,6 and 7). The number of extracted relations depends slightly on the similarity measure type. Low minimum confidence will lead to excessive numbers of related concepts, thus resulting increases in calculation complexity and processing time. High minimum confidence will lead to discard important related concepts, thus resulting in the concept context being unable to represent concept. One important consideration for concept context construction is therefore on how to define appropriate minimum confidence. In this study, experiments were used to analyze the impacts of different min confidence on concept context construction.

To validate the proposed method, we performed experiments in different domains over Al-Hdith corpus. Al-Hadith corpus contains 7397 hadiths and divided into 97 books under each book different chapters.

5.2 The Experimental Result

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. Because the lack of gold standard, it is difficult to verify the performance of relation extraction methods. For example, to compute the recall of a relation extraction method, we need to know all valid relations of the domain. In our experiments, we use the precision to evaluate the performance of the proposed method:

$$precision = \frac{number\ of\ correctly\ retrieved\ relations}{number\ of\ relations\ retrieved} \qquad (11)$$

We also, ask an expert to rate the extracted relations between concepts as Highly related (Hr), Related (R) and Not Related (NR). Then, we compute the average score for each similarity measure. Tables 8 and 9 show the evaluation results when the minimum confidence =0.25 and 0.50 respectively. Figures 6 and 7 show the comparison between measures at the confidence =0.25 and 0.50 respectively.

From Figure 6, we found that Lin measure has the highest precision 83% at confidence =0.25. As it can be seen in Figure 7, the results of the evaluation seem promising. It is worth noting, for instance, that the method obtains a precision value of 85% at minimum confidence = .50 for Jaccard and Dice measures.

Figures 6 and 7 showed that the Jaccard and Dice similarity measures extracted a high related concepts outperforms the Cosine similarity. For example, seed concept = تعرق العضد , concept = الله with similarity 0.23801, is not related concepts extracted by Cosine similarity while discarded by Jaccard. Also, highly related concepts for example, seed concept = الكتف و الجنب, concepts = شاه with similarity 0.31192, extracted by Jaccard and discarded by Cosine similarity. From Tables 3,4,5 and 6, we observed that the relations extracted from one measure complement another.

To see how our method performs in different domains, we compared the performance of our method on medicine domain and food domain. (see Figures 10,11,12 and 13). The first thing to notice is that in all similarity measures, the best results are obtained at min-conf = .50 for Jaccard and Dice. Moreover, we can observe that precision of similarity measures tends to decrease while the minimum confidence decreases. This evaluation also shows that Lin measure outperforms Cosine similarity measure.

	Hr	R	Nr	Precision
Cosine	0.58	0.18	0.22	0.77
Jaccard	0.59	0.20	0.19	0.79
Dice	0.59	0.2	0.19	0.79
Lin	0.60	0.22	0.18	0.83

TABLE 8: Evaluation results for Medicine domain at min-Conf=0.25.

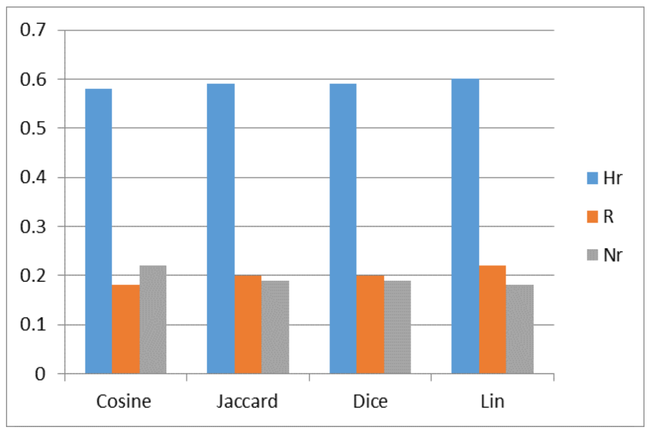

FIGURE 6: Comparison of similarity measures at min-conf=.25 for Medicine domain.

	Hr	R	Nr	Precision
Cosine	0.63	0.20	0.16	0.83
Jaccard	0.65	0.19	0.13	0.85
Dice	0.65	0.20	0.15	0.85
Lin	0.61	0.21	0.15	0.81

TABLE 9: The evaluation results for Medicine domain at min-conf=.50.

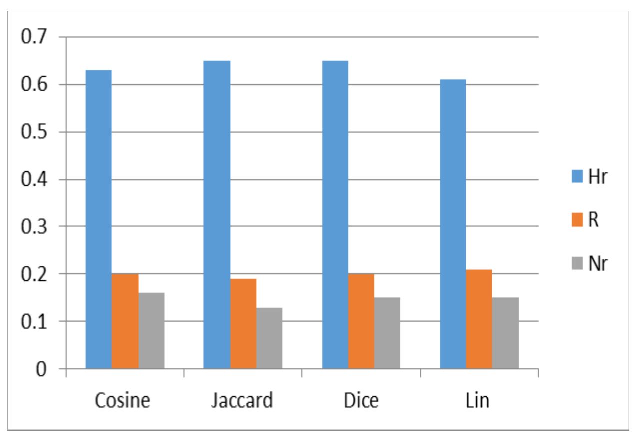

FIGURE 7: Comparison of similarity measures at min-conf=.50 for Medicine domain.

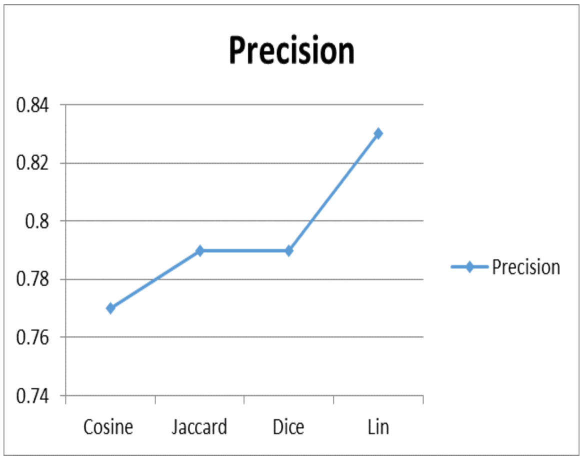

FIGURE 8: The precision at min-conf=.25 for Medicine domain.

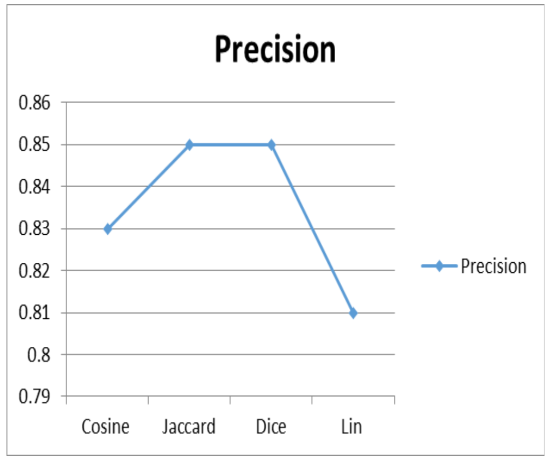

FIGURE 9: The precision at min-conf=.50 for Medicine domain.

With the above results, we can see from Tables 10 and 11 the precision is range from 71% to 72% at minimum confidence =0.25 and from 75% to 79% at minimum confidence =0.50 in Food domain. Such a result demonstrates the effectiveness of our method for relation extraction in different domains. But comparing with the Medicine domain, see Tables 8 and 9, we find the precision is range from 77% to 83% at minimum confidence =0.25 and from 81% to 85% at minimum confidence =0. 50.

The results (see Figures 14 and 15) show that our method for relation extraction performs better in Medicine domain than in food domain. As we explained in concept extraction method, some errors were created due to errors of the subsequent POS tagging and noun phrase extraction. In Medicine field, the noun phrases are more than in food field. While food domain has more verbs. Accordingly, when construct context for concepts a lot of knowledge discarded because it is in the verb form.

	Hr	**R**	**Nr**	**Precision**
Cosine	0.54	0.15	0.30	0.70
Jaccard	0.57	0.13	0.28	0.71
Dice	0.49	0.22	0.27	0.72
Lin	0.49	0.23	0.27	0.72

TABLE 10: The evaluation results for Food domain at min-conf=0.25.

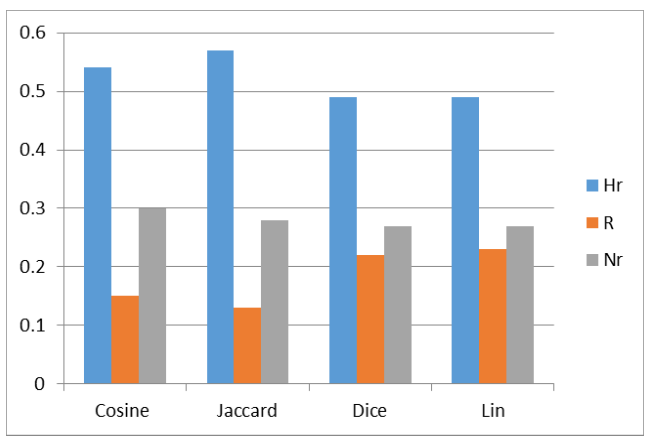

FIGURE 10: Comparison of similarity measures at min-conf=.25 for Food domain.

	Hr	R	Nr	Precision
Cosine	0.49	0.26	0.25	0.75
Jaccard	0.57	0.22	0.20	0.79
Dice	0.52	0.26	0.20	0.79
Lin	0.55	0.21	0.22	0.77

TABLE 11: The evaluation results for Food domain at min-conf=0.50.

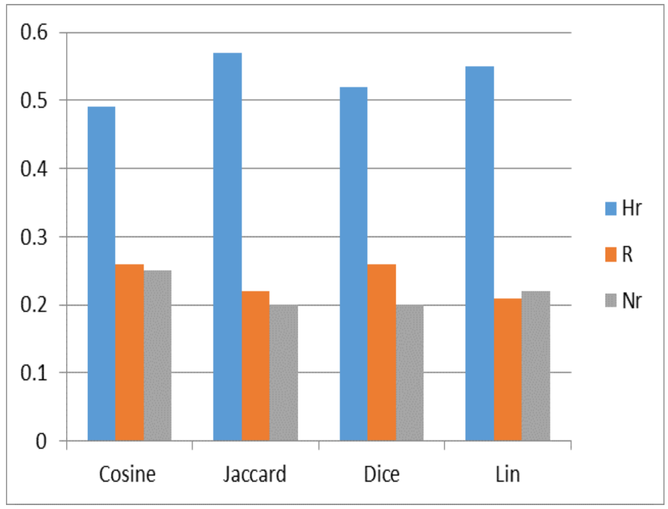

FIGURE 11: Comparison of similarity measures at min-conf=.50 for Food domain.

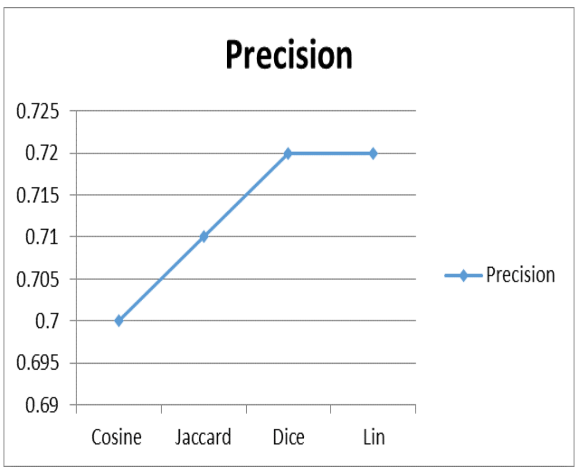

FIGURE 12: The precision at min-conf=.25 for Food domain.

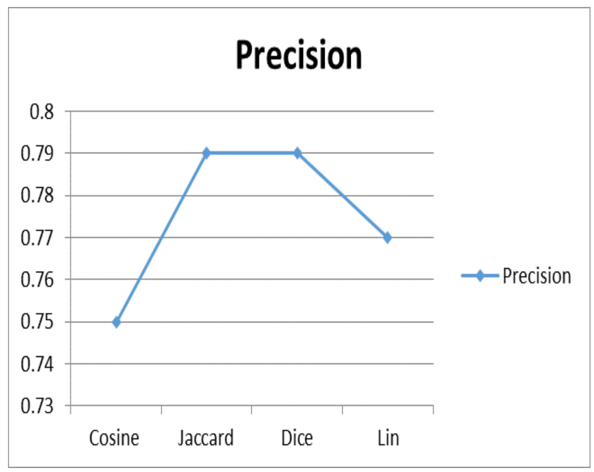

FIGURE 13: The precision at min-conf=.50 for Food domain.

	Cosine	Jaccard	Dice	Lin
Medicine	0.77	0.79	0.79	0.83
Food	0.70	0.71	0.72	0.72

TABLE 12: The Precision results for Medicine and Food domain at min-conf=0.25.

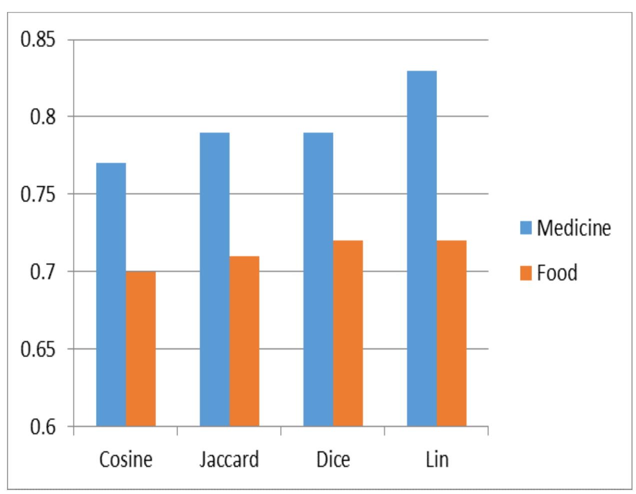

FIGURE 14: The performance of algorithms for relation extraction in Medicine domain and Food domain at min-conf= 0.25.

	Cosine	Jaccard	Dice	Lin
Medicine	0.83	0.85	0.85	0.81
Food	0.75	0.79	0.79	0.77

TABLE 13: The Precision results for Medicine and Food domain at min-conf=0. 50.

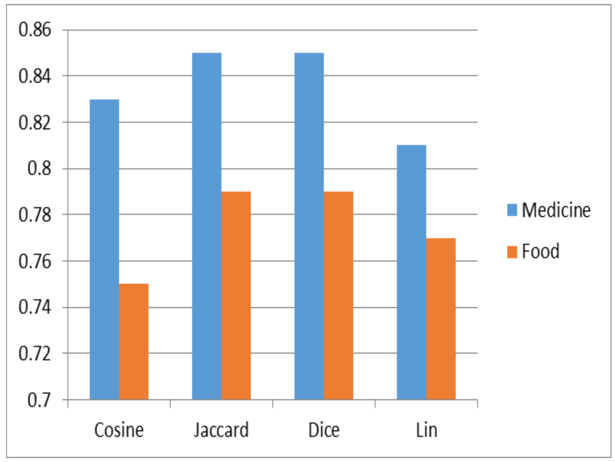

FIGURE15: The performance of algorithms for relation extraction in Medicine domain and Food domain at min-conf= 0.50.

6. CONCLUSION

In this paper, we have presented a new method for relation extraction from Arabic texts. It uses a new context definition of concepts based on relevance analysis and semantic similarity measures to extract relations from domain corpus. Also, we constructed an initial taxonomy using a seed concepts and noun-phrase based patterns.

Our approach has several advantages over other methods discussed in the previous section, the most important of which is the ability to handle implicit relation extraction. And it performs in unsupervised manner which means, no need for training data. Also it is used effectively for relation extraction in various domains. In order to solve the over generation problem of co-occurrence, we proposed the relevance measure that measure the degree of association between concepts. In this study, experiments were used to analyze the impacts of different minimum confidence on concept context construction.

Compared to other methods described in the previous section for Arabic relation extraction based on Hearst' patterns, our result outperforms the others in term of precision. The Precision on Holy Quran set reached 76.28% for [3]. And no quantitative results were provided to compare our work to the method using association rule for hadith [7]. [11] Experimented with the Hadith corpus and recorded 60% as the most decreased rate while the co-occurrence method reached 67% as the best result.

From the results observed in the analyses performed we conclude that:

Our experiments support our assumption about the usefulness of our method for relation extraction. As shown through the evaluation, our method has a strong ability to extract implicit relations. This overcomes the problems that have been found in the methods that are based only on the Hearst patterns that extract explicit relations only. And Hearst approach works well only for the documents that contain a lot of patterns. From our study to the corpus characteristics, we found that the relations between concepts implicit and not explicitly described by patterns.

From our results, we have determined that using the definition of concept context based on relevance analysis and similarity measures increases the precision of the relation extraction method. From the output of the method, we found that most of the relations in the domain are extracted by our method, which constitute a domain knowledge that needed to the expert to construct ontology. That means the recall is high. However, we did not compute it because a reference standard for the domain (all relations in the domain) is not available.

We observed that, the relations extracted from one measure complement another. So, a hybrid similarity measure is proposed.

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 most of the relations in food domain expressed using the verbs. On the other hand, most of the important domain relations are extracted correctly. The results show the high effectiveness of the proposed approach to extract relations for Arabic ontology construction. For the future work, we will continue to evaluate and compare results of other domains and we will propose a hybrid similarity measure.

7. REFERENCES

[1] P. Cimiano." Ontology Learning and Population from Text: Algorithms, Evaluation and Applications", in Studies in Philosophy and Religion, Springer, 2006.

[2] A.Al-Arfaj and A. Al-Salman. (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), 2014, pp. 737-742.

[3] M. AL-Zamil and G. Al-Radaideh. “Automatic Extraction of Ontological Relations from Arabic Text.” Journal of King Saud University - Computer and Information Sciences,26(4),pp. 462–472, 2014.

[4] M. Al-Yahya., L. Aldhubayi and S. Al-Malak. “A Pattern-Based Approach to Semantic Relation Extraction Using a Seed Ontology,” in Proceeding of IEEE International Conference on Semantic Computing, Newport Beach, California, 2014, pp,96-99.

[5] I.Sarhan., Y. El-Sonbaty and M. El-Nasr. ‘Semi-Supervised Pattern-Based Algorithm for Arabic Relation Extraction”. In proceeding of IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), San Jose, CA, USA, 2016 pp. 177-183.

[6] A.Alotayq. “Extracting relations between Arabic named entities”, Lecture Notes in Computer Science, vol 8082, Springer-Verlag, Berlin Heidelberg, Pilsen, pp.265-271, 2013.

[7] F. Harrag., A. Alothaim., A. Abanmy., F. Alomaigan and S. Alsalehi. " Ontology Extraction Approach for Prophetic Narration (Hadith) using Association Rules.. International Journal on Islamic Applications in Computer Science And Technology, 1(2), pp. 48-57, 2013.

[8] S. El-salam., E. El Houby., A. Al Sammak and T. El-Shishtawy. “ Extracting Arabic relations from the web”. International Journal of Computer Science & Information Technology (IJCSIT), vol 8, no 1,pp. 85-102, 2016.

[9] N. Taghizadeha., H. Failia,H and J. Malekib. “Cross-Language Learning for Arabic Relation Extraction”., Procedia Computer Science, 142, pp. 190–197 . 2018.

[10] I.Bounhas , B. Elayeb ., F. Evrard and Y. Slimani. “ArabOnto: Experimenting a new distributional approach for building Arabic ontological resources.” International Journal of Metadata, Semantics and Ontologies (IJMSO), 6(2), pp. 81 – 95, 2011.

[11] W. Lahbib., I. Bounhas., B. Elayeb., F. Evrard and Y. Slimani. “A Hybrid Approach for Arabic Semantic Relation Extraction,” in Proceeding of The 26th International Florida Artificial Intelligence Research Society (FLAIRS), 2013,pp. 315-320.

[12] H. Haddad., J. Chevallet and M. Bruandet. “Relations between Terms Discovered by Association Rules,” in Proceeding of 4th European Conference on Principles and Practices

of Knowledge Discovery in Database (PKDD2000), Workshop on Machine Learning and Textual Information Access, Lyon France, 2000.

[13] J.R. Curran and M. Moens. "Improvements in automatic thesaurus extraction," in Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition, 2002, pp. 59-66.

[14] D. Lin., S. Zhao., L. Qin and M. Zhou. "Identifying synonyms among distributionally similar words," in Proceeding of 18th International Joint Conference on Artificial Intelligence (IJCAI), 2003, pp. 1492-1493.

[15] O. Ferret. " Testing semantic similarity measures for extracting synonyms from a corpus," in Proceeding of 7th Conference on International Language Resources and Evaluation (LREC), Valeeta, Malta, 2010, PP. 3338-3343.

[16] L. Meng., R. Huang and J. Gu. "A Review of Semantic Similarity Measures in WordNet." International Journal of Hybrid Information Technology. 6(1), pp. 1-12, 2013.