JOURNAL OF ENGINEERING RESEARCH AND TECHNOLOGY, VOLUME 9, ISSUE 2, OCTOBER 2022

Received on (01-02-2022) Accepted on (19-04-2022)

Exploiting Wikipedia to Measure the Semantic Relatedness between Arabic Terms

Basel AlHaj and Iyad AlAgha

https://doi.org/10.33976/JERT.9.2/2022/1

Abstract—Measuring the semantic relatedness between words or terms plays an important role in many domains such as linguistics and artificial intelligence. Although this topic has been widely explored in the literature, most efforts focused on the English text, while little has been done to measure the similarity between Arabic terms. A growing number of semantic relatedness measures have relied on an underlying background knowledge such as Wikipedia. They often map terms to Wikipedia concepts, and then use the content or hyperlink structure of the corresponding Wikipedia articles to estimate the similarity between terms. However, existing approaches mostly focused on the English version of Wikipedia, while limited work has been done on the Arabic version. This work proposes an approach that takes advantage of Wikipedia features to measure the relationship between Arabic terms. It exploits two types of relations to gain rich features for the similarity measure, which are: the context-based relation and the category-based relation. The context-based relation is measured based on the intersection between incoming links of Wikipedia articles, while the category-based relation is measured by utilizing the taxonomy of Wikipedia categories. The proposed approach was evaluated based on a translated version of the WordSimilarity-353 benchmark dataset. The results show that our approach generally outperforms several approaches in the literature that use the same dataset in English. However, the poor structure and content of the Arabic version of Wikipedia compared to the English version has resulted in several incorrect similarity scores.

Index Terms—Semantic relatedness, Arabic text, Wikipedia, Text Similarity.

I Introduction

Measuring semantic relatedness between terms is an important issue in natural language processing, and is used in many application areas such as information extraction and retrieval, text summarization, document classification and clustering, and question answering. Terms can be related either lexically or semantically: Terms have a lexical relevance if they have a similar sequence of characters such as 'underestimate' and 'understand' or 'withhold' and 'withdraw'. Terms have semantic relatedness when they are frequently used in the same context. For example, "job" and "money" are semantically related even though they are not lexically related [1, 2].

Lexical relatedness is often calculated by using String-Based Algorithms (SBA), which are based on string sequences and character composition to determine if two strings are similar or not. Semantic relatedness is often measured using Corpus-Based (CBA) and Knowledge-Based algorithms (KBA). CBA is based on information gained from large corpora to calculate similarity between terms, while KBA uses information derived from semantic networks [2]. Humans depend on a huge amount of background knowledge in analyzing the meanings of terms. Therefore, any automated attempt to calculate the semantic relatedness between terms should be based on external sources of knowledge [3].

Wikipedia today represents one of the largest sources of knowledge and covers a large number of knowledge domains. Due to this importance and popularity, many works have exploited Wikipedia as a knowledge source to measure the semantic relatedness between terms. Some of these works have exploited the structure of Wikipedia articles such as categories, hyperlinks, and templates [4, 5]. Another line of works has attempted to measure similarity based on the natural language processing of the textual content of Wikipedia articles [6-9]. The aforementioned works have focused solely on English text and used the English version of Wikipedia. To the best of our knowledge, few efforts have benefited from the Arabic version of Wikipedia to improve the contextual understanding of the Arabic text.

In this work, we propose a Wikipedia-based approach for measuring the semantic relatedness between Arabic terms. Wikipedia is particularly selected as a knowledge source for our approach due to its large content and wide coverage of different domains of knowledge. This enables our approach to measure similarity between terms from different domains of knowledge. Given any two Arabic terms, our approach should give a score that indicates the degree of similarity between them. The proposed approach exploits the hyperlinks between Wikipedia articles and the taxonomy of categories to capture the semantic similarity between terms. The hyperlink structure is used to determine the context-based relation between the articles that correspond to input terms.

Wikipedia categories are also used to group articles with similar or related subjects together. The category graph of the Arabic version of Wikipedia is constructed and analyzed to compute the category-based relation between terms.

The contributions of this work can be summarized as the following: 1) It presents an approach to measure the semantic relatedness between Arabic terms using the Wikipedia' hypertext structure and category graph. We provide the source code of the proposed approach through the following link: (https://github.com/BaselAlhaj/SemanticRelatedness). 2) By comparing our approach with similar approaches that use the English version of Wikipedia, we can assess the reliability of the Arabic version of Wikipedia as compared to the English version and inform the research community of the potential of Arabic Wikipedia for measuring relatedness between Arabic terms. 3) It provides a hands-on-experience in processing Wikipedia content to enable searching in and mapping to Wikipedia articles, as well as the construction of category graph to measure category-based relation. We believe that this will be of importance to practitioners and researchers who are interested in exploiting the Arabic version of Wikipedia.

II Related Works

A variety of semantic similarity methods have been proposed in the literature, which can be generally classified into three main categories [10, 11]: 1) knowledge-based methods, and 2) corpus-based methods, 3) Deep learning methods. In what follows, we discuss these categories of similarity methods, and then review the related works on Arabic text similarity measures.

A. Knowledge-based Semantic Similarity Methods

Knowledge based semantic similarity methods estimate the similarity between terms based on the information extracted from background knowledge sources. These methods often rely on the structured representation offered by the background knowledge [12]. This structure often comes as a set of concepts connected with relations. Examples of knowledge sources widely used for similarity methods include WordNet, Wiktionary, Wikipedia, and BabelNet [13].

Knowledge-based similarity methods can be classified according to the underlying principle into four categories [10]: edge-counting methods, feature-based methods,, content-based methods and hybrid methods. Edge-based methods consider the underlying knowledge as a graph connecting concepts taxonomically and count the edges between terms to measure the similarity. The greater the distance between the terms, the less similar they are. In general, the limitation of edge counting methods is that the distance often fails to capture the similarity between terms.

Feature-based methods calculate similarity as a function of properties of the terms based on the neighboring terms, or the different meanings of the term in the glossary [12]. For example, the Lesk measure [14] estimates the relatedness between two terms based on the overlap of their meanings in a background dictionary like WordNet. Jiang et al. [15] proposed an approach that measures the semantic similarity using the glosses of concepts present in Wikipedia. The main problem with feature-based methods is their dependency on the existence of semantic features, which are not always present in the background knowledge.

Information content-based methods attempt to estimate the similarity between terms based on what is called the Information Content (IC) of the term. The IC of a term is defined as the information derived from the context where the term appears in. A high IC value indicates that the term is more specific and describes a topic with less ambiguity, while a lower IC means that the term is more abstract in meaning [16]. A numerous number of extensions have been proposed to measure the term's IC by exploiting different features of the underlying structure of the background knowledge [17-19]. In this work, we use and evaluate two methods to calculate the IC value of a term based on the taxonomy of concepts in the Arabic Wikipedia.

Hybrid knowledge-based methods combine various measures from the three categories aforementioned to better capture the similarity between terms. For example, Goa et al. [20] proposed a method that uses three different strategies that include the depths of all the terms in WordNet along with the path between the two terms, the depth of the least common subsumer of the terms, and the IC value of the terms. In general, knowledge-based measures are highly dependent on the richness, divergence, and recentness of the underlying knowledge.

B. Corpus-based Similarity Methods

Corpus-based methods calculate the semantic similarity between terms using the information retrieved from large corpora. There is a wide variety of corpus-based techniques for measuring the semantic similarity between texts. Latent Semantic Analysis (LSA) is one of the most popular and widely used corpus-based methods [21]. It is a statistical text analytics method that can uncover the conceptual content within unstructured data by using Singular Value Decomposition (SVD). Several works have used LSA to measure similarity between terms [22, 23]. Hyperspace Analogue to Language (HAL) is another corpus-based method that captures the statistical dependencies between terms by considering their co-occurrences in a surrounding window of text [24, 25]. Word-Alignment models present another line of corpus-based methods that calculate the semantic similarity between sentences based on their alignment over a large corpus [26, 27]. Latent Dirichlet Allocation (LDA) [28] is another technique that is widely used for topic modeling tasks, and it has the advantage of reduced dimensionality [29]. Normalised Google Distance (NGD) is another corpus-based measure of semantic similarity that is derived from the Google search engine. It is based on the assumption that two words are highly related if they occur together frequently on web pages [30]. Word-attention models [31, 32] are of

the most recent and promising corpus-based methods that differ from traditional semantic similarity methods in that they can capture the importance of words from underlying corpora before calculating the semantic similarity.

C. Deep Learning Methods

Semantic similarity methods have recently exploited recent developments in neural networks and deep learning techniques to enhance performance. Plenty of works proposed methods to measure semantic similarity between terms by using Convolutional Neural Networks (CNN) [33, 34], Long Short Term Memory (LSTM) [35], Bidirectional Long Short Term Memory (Bi-LSTM)[36], Recursive Tree LSTM [37], and Transformers [38]. Despite the great potential of deep learning methods, their main limitation is that they are computationally expensive and require large training sets to work effectively. In contrast, the approach proposed in this work is unsupervised, and thus does not require labelled data.

D. Arabic Semantic Similarity Methods

In the context of Arabic text, several works proposed methods to measure the similarity between documents [39, 40], sentences [41-43] and words [44, 45]. However, most of these methods are based on word co-occurrences or word embeddings, which help more for capturing syntactic rather than semantic similarity between words. Few methods tackled semantic techniques for Arabic text similarity. For example, Almarsoomi, et al. [44] used the measure proposed by Li et al. [46] to calculate the similarity between words by exploiting different attributes from the Arabic WordNet. Froud et al. [45] measured the semantic similarity between two words by using the Latent Semantic Analysis (LSA) model and demonstrated the difference between using stemming and light stemming in the preprocessing phase.

Recently, an increasing number of works have exploited the Arabic version of Wikipedia for different purposes in computer science. Some works use the structured-content of Wikipedia to construct ontologies [47, 48]. Others used Wikipedia features and hyperlink structure to build Arabic-named entity corpora [49, 50] or for entity linking[50]. Wikipedia-based categories have been also used to support the classification of Arabic text [51], the open-domain text tagging [23], and the search query expansion [52]. This work adds to the previous knowledge by leveraging the Arabic Wikipedia to measure semantic relatedness between Arabic terms.

III Overview of the Proposed Approach

In general, our approach exploits the structure of Wikipedia to measure two types of relations between terms, which are the context-based relation and the category-based relation. The context-based relation estimates the relatedness between two terms based on the commonality between the corresponding articles in Wikipedia. In the context of Wikipedia, any two terms can be related if the corresponding articles share common links. In our approach, incoming links from articles are used to compute relatedness. The more incoming links shared between articles, the more related they are.

The category-based relation depends on the categories that are used to classify Wikipedia articles. Wikipedia articles are categorized by using a taxonomy of predefined categories. If articles share same or related categories, via a child-parent relation for example, then these articles are likely to be related.

Our approach combines both category-based and context-based relations to estimate the relatedness between any two Wikipedia articles. intuitively, the relatedness between the articles denotes the relation between the terms representing them. Figure 1 depicts our approach for measuring the semantic relatedness between sample input terms A and B. The first step is to match the terms to the Wikipedia articles that best describe them. Then, both the context-based and the category-based relations between the two articles are measured. Several computations are performed at this phase to analyze the hyperlink and category structures. The final relatedness score will be the average of the two relation scores mentioned above.

IV Context-based Relation

The context-based relation between terms reflects how often these terms share contexts. Wikipedia articles contain many hyperlinks that refer to other articles. In our approach, we depend on incoming links to represent shared contexts between two articles. The greater the number of shared incoming links, the higher the context-based relation is. The context-based relation between them two Wik-

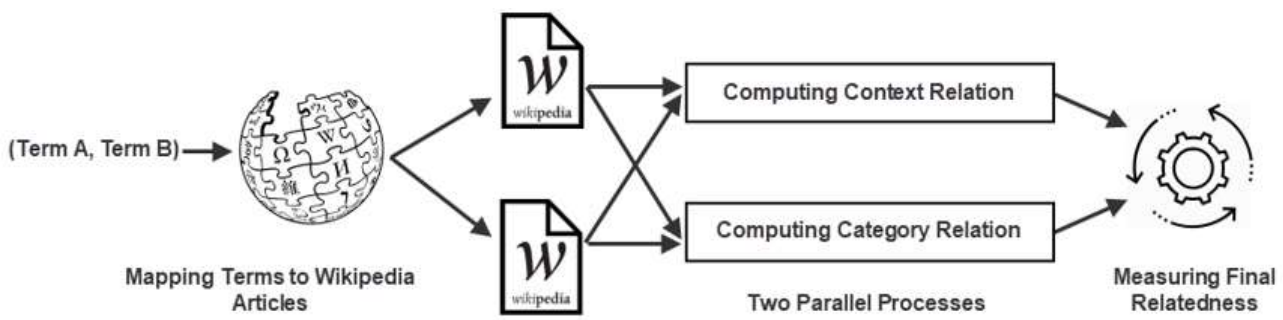

Figure 1. Measuring the semantic relatedness between terms by exploiting context-based and category-based relations

ipedia articles can be measured using the following equation from [3]:

$$context\_rel(a,b) = 1 - \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log|N| - \log(\min(|A|,|B|))} \quad (1)$$

where a and b are any two articles from Wikipedia, A and B are the sets of incoming links to a and b respectively. N is the total number of articles in Wikipedia.

V CATEGORY-BASED RELATION

Wikipedia provides many categories that are used in each article to determine its scope. Articles belonging to the same Wikipedia category are related. These categories are used in our approach to determine the relatedness between Wikipedia articles as shown in Figure 2. For any two articles a and b, let S_1= {c_{11}, c_{12}, ..., c_{1n}} and S_2= {c_{21}, c_{22}, ..., c_{2m}} be the sets of categories that a and b belong to respectively. n and m are the sizes of S_1 and S_2 respectively. In our approach, the pairwise relation between every two categories (c_{1i}, c_{2j}) is calculated, where $c_{1i} \in S_1$ and $c_{2j} \in S_2$. Then, the overall relation between S_1 and S_2 is calculated by combining pairwise relation scores.

First, the relation between c1i and c2j is calculated using the following equation.

$$pairwise\_cat\_rel(c_i, c_j) = \frac{\mathrm{IV}(\mathrm{LCA}(c_i,c_j))}{\mathrm{IV}(c_i)+\mathrm{IV}(c_j)} \quad (2)$$

where $pairwise\_cat\_rel(c_i, c_j)$ is the category-based relation between c_i and c_j, $LCA(c_i, c_j)$ is the lowest common ancestor of c_i and c_j, and IV(c) is the information value of the category c. The calculation of LCA(c_i, c_j) and IV(c) is explained in the subsequent sections.

for each $c_{1i} \in S_1$, we find best(c_{1i}), which is the maximal pairwise similarity between c_{1i} and any category c_{2j} in S_2. Similarly, we find $best(c_{2j})$ for each $c_{2j} \in S_2$. The overall category-based relation between S_1 and S_2 is calculated using the following equation [53]:

$$cat\_rel(S_1, S_2) = 0.5 * \frac{\sum_{i=1}^{n} best(c_{1i})}{n} + 0.5 * \frac{\sum_{j=1}^{m} best(c_{2j})}{m} \quad (3)$$

Figure 2 summarizes the process of calculating the category-based relation.

VI INFORMATION VALUE

As shown in Equation 2, measuring the category-based relation is based on the information value of categories. The information value indicates the specificity of the category. For example, the category "هندسة البرمجيات" is more specific that the general category "علم الحاسوب". Thus, the former category contains more information value than the latter category when both are assigned to a single article. The general categories, or top-level categories, are not reliable in measuring relatedness because they often do not help distinguish between articles unlike the specific categories. Thus, we need to give each category an information value based on its specificity.

To determine the specificity of a category, the Wikipedia category graph is first constructed. Then, two measures are used to calculate the information value for a category. These measures use: 1) the depth of the category in the Wikipedia category graph. 2) the number of descendants of the category. These measures are explained as follows:

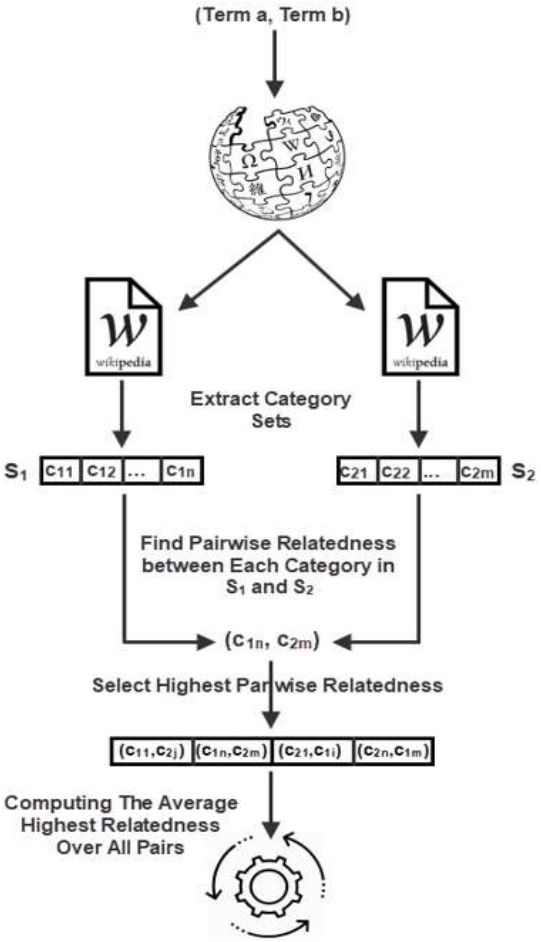

Figure 2. Measuring the category-based relation between terms a and b

The first measure depends on the depth of a given category in the Wikipedia category graph to determine the appropriate information value for it by using the following equation.

$$catDepthIV(c) = \frac{\log(\mathrm{maxDepth(c)})}{\log(graphMaxDepth)} \quad (4)$$

Where $catDepthIV(c)$ is the information value of the category$c$. The maxDepth(c) is the depth of c in the entire Wikipedia category graph, and $graphMaxDepth$ is the maximum depth of Wikipedia category graph. This measure indicates that the larger the depth of c, the larger information value it has. In general, top-level categories are of low depth, and are often general and contain less information value compared to high-depth categories. This measure was inspired by existing research on the specificity of graph nodes such as [54] and [55].

The second measure uses the descendants (subcategories) of a category to determine the appropriate information value for it as follows.

$$catDescendantsIV(c) = 1 - \frac{\log(\text{des(c)})}{\log(N)} \quad (5)$$

Where $catDescendantsIV(c)$ is the information value of category c, des(c) is the number of descendants of category c and N is the set of all categories in Wikipedia. It is assumed that the category with a large number of descendants is more general and thus has less information value. In contrast, categories with few or no descendants are likely to be more specific and thus have more information value. This metric was also inspired by existing works that present metrics for graph-based similarity such as [54].

Given the above two measures of the category's information value: $catDepthIV$ and $catDescendantsIV$, only one measure will be used to calculate information values of categories in Equation 2. Part of the experiments that we conducted in the evaluation aimed to examine the two measures of information value in order to choose the best of them to be used in Equation 2.

VII Combined Relatedness Measure

In the above sections we showed how to measure the context-based and category-based relations between any pair of terms respectively. The overall relatedness between two terms is then calculated as the average of the category-based relation and the context-based relation values using the following equation.

$$Sim.relatedness(a,b) = \frac{context\_rel(a,b) + cat\_rel(a,b)}{2} \quad (6)$$

I Implementation Highlights

After formally presenting our approach, the following sections provide a step-by-step guide on the implementation details including the processing of Wikipedia content, the mapping of the terms to Wikipedia articles, and the construction of Wikipedia category graph. We also show how we handled the challenges that can be encountered when processing the Wikipedia graph, such as graph cycles and creating the category depth and descendants' maps.

Figure 3 shows the components of our implementation of the proposed approach. Part 1 in Figure depicts the preprocessing of the Wikipedia dump file to store its content in a local database, and this process is performed once at the beginning of the work. Part 2 shows the Article Matcher module that is used to match the input terms to the corresponding Wikipedia articles. Part 3 represents the process the constructing of category graph, the map of the descendants of the category, and the map of the depth of the category. Given two input terms, the process starts by mapping these terms to the corresponding Wikipedia articles using the Article Matcher module; then the semantic relatedness score is calculated by calculating the context-based and category-based relations and averaging them as in Equation 6. The modules in Figure 3 are explained in detail in what follows:

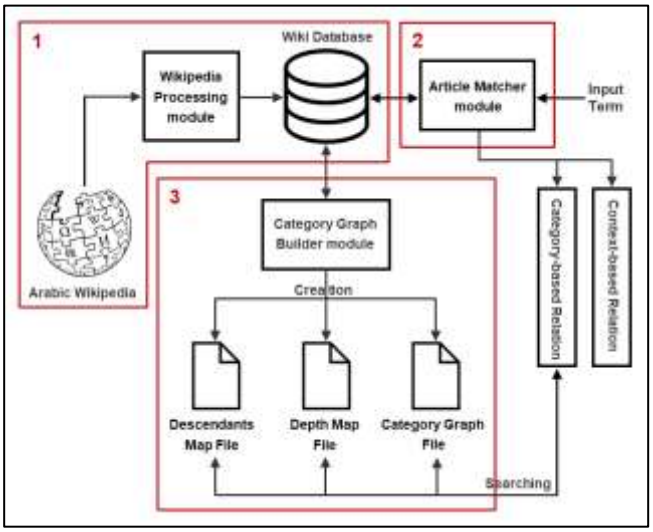

Figure 3. the components of our implementation of the proposed approach for measuring semantic relatedness between Arabic terms.

A. Wikipedia Processing Module

To access the Arabic Wikipedia and extract the required information, we downloaded the XML Dump file of the Arabic Wikipedia, on the 1st Feb. 2021. The information about the downloaded dump file is shown in Table 1.

TABLE 1
Information about the downloaded dump

XML Dump File Size	736 MB
Size after extraction	4.87 GB
Number of All Pages	1243905
Number of Content pages (Articles)	648399
Number of Redirect Pages	584630
Number of Disambiguation Pages	16663
Number of Categories	560154
Number of Articles used in our work	515094
Number of Categories used in our work	547396

After that, the XML dump file parsing process was performed, and the information was extracted and stored in a local database. The database contains tables for pages, page-inlinks, page-outlinks, page-redirects, page-categories, page-mapline, metadata, categories, category-inlinks, category-outlinks, and category-pages. Access to Wikipedia information during work will be done by querying the database. We used JWPL (Java Wikipedia Library) [56] to parse the Wikipedia dump file. JWPL is a free, Java-based application programming interface, which allows a structured access to all information in Wikipedia.

After populating the tables in the database, we found some articles that do not have incoming links. The incoming links are important in our work since one of the metrics used depends on the number of incoming links to articles. Therefore, these articles have been discarded and deleted from the database, knowing that the total number of content pages with no incoming links is 133305 articles. In addition, some Wikipedia categories that are used for editing and managing articles, known as administrative categories, were discarded as they negatively affect the evaluation of the semantic relatedness. Examples of administrative categories include "ويكيبيديا", "مشاريع ويكي", "مقالات", "قوالب", "قالب", "بذرة" and "صناديق المعلومات".

B. Article Matcher Module

Article matcher (see Figure 3) is a component responsible for mapping each input term to the corresponding Wikipedia article. The goal is to obtain the appropriate Wikipedia article to be used to measure semantic relatedness. For each term, the matcher first converts the text of the term to a normalized form. The normalized text is used to check whether there is an article matching the term in the database or not, and will improve the result of the matching process. For example, the term "اللغة العربية" will be converted to the following normalized format: "(ا|إ|أ|آ)لعربي(ه|ة)_(ا|إ|أ|آ)للغ(ه|ة)", where suffixes and prefixes will be separated from the original term. In addition, different forms of Arabic letters will be considered when matching with Wikipedia article. In addition, redirect pages should be excluded from the matching process because they have no content or categories, and thus could impede the semantic relatedness score.

C. Handling Term Disambiguation

Some terms are ambiguous in the sense that they can have multiple meanings. Wikipedia provides disambiguation articles for these terms, whereas each disambiguation page contains a list of possible senses for the term. For example, the term "عين" in Wikipedia is a redirect page to the disambiguation page "عين (توضيح)", which contains a list of articles with different meanings such as "عين (طب)", "عين (حرف)" and "عين (ماء)". When a disambiguation page is retrieved from the Article Macher for an input term, the list of all senses is considered when the relatedness with the other input term is calculated. The sense that achieves the highest relatedness score is used, and the other senses are skipped. For example, if the input terms to our approach are "عين" and "نبع". Since the first term "عين" has three senses as explained above, the semantic relatedness between ease sense and the term "نبع" is measured, and the sense that gives the best relatedness score will be used.

D. Category Graph Builder Module

The measurement of category-based relations depends essentially on the calculation of the information value for each category, as explained before. To determine the information value, two structures should be constructed, which are: the category depth map and the category descendants map. A category graph to speed up the processing of categories and calculation of results.

In order to construct the category graph, a directed graph was created, and Wikipedia categories are added to the graph as vertices. To determine the edges of the graph, the incoming links (parent categories) and outgoing links (children categories) of each category are used. For a category (c) that has the set of incoming links IN_LINKS={ic1, ic2, ..., icn} and the set of outgoing links OUT_LINKS={oc1, oc2, ..., ocm}, a graph edge is created from each ici∈ IN_LINKS to each ocj∈ OUT_LINKS. Finally, we get a graph whose vertices are Wikipedia categories, and edges are the links between these categories. Creating a graph consumes a lot of time so we created it once and saved it as a serializable object in a file. When needed, it is loaded from the file rather than recreating it.

The directed graph should not contain any self-directed edge such as the one shown in Figure 4. Self-directed edges cause infinite loops during depth computation. These edges can be easily detected and ignored by finding the intersection between the incoming links and outgoing links, that is: cycles = IN_LINKS ∩ OUT_LINKS.

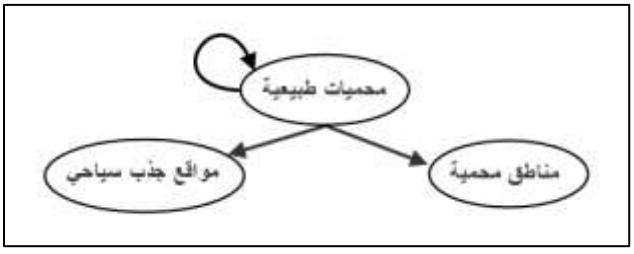

Figure 4. Self-directed edge

In addition, cycles as the one shown in Figure 5 should also be eliminated because they can cause infinite loops when calculating the depth of the category graph as required in Equation 4. To find cycles, a first-depth-first (FDS) traversal was performed starting from top-level categories, and each visited vertex is marked. If the vertex is visited twice during the FDS, the incoming edge through which the vertex is reached for the second time is removed. Our experiments showed that this strategy has successfully eliminated most cycles in the constructed category graph.

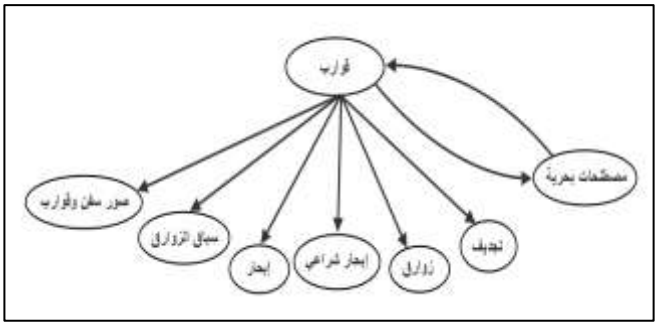

Figure 5. A graph cycle

E. Depth and Descendant Maps

The estimation of the information value of a category requires extracting some information from the category graph, such as the depth of a category and the number of category descendants (refer to Equations 4 and 5). Knowing that extracting this information for each category at run time takes a lot of time, it would be better to extract them one time and store them to be used when needed. For this purpose, two maps were created: VertexDepthMap and DescendantsMap. VertexDepthMap is used to store the path from each category in the graph to the root, while DescendantsMap is used to store the number of descendants of each category in graph.

To create VertexDepthMap, the leaf vertices of the graph are first extracted, which are the vertices that have no outgoing edges. For each leaf vertex, all up-level vertices through the incoming edges are extracted, and this is performed recursively until reaching the root. During the recursion process, the different paths from each vertex to the root are compared, and the longest path is retrieved. DescendantsMap creation also begins from the leaf vertices of the graph. While moving up from the leaf vertices, the children for each vertex are counted and stored in map. Whenever moving to an upper level in the graph, the number of child vertices is calculated in the same way as before, and the number of descendant vertices is added by being retrieved from map. Moving to a higher level is continued until reaching the root.

F. User Interface

We provide a simple user interface, as shown in Figure 6, that allows the user to input two Arabic terms and get the relatedness score between them. The user should choose the appropriate settings and click the button "Compute Relatedness". The value of the semantic relatedness is displayed in a range from 0 to 1, where 0 denotes no relatedness, and 1 denotes maximum relatedness.

Figure 6. User interface to compute the semantic relatedness between input terms

II EVALUATION

Similar approaches from the literature have been often evaluated by being compared with other approaches [3, 57, 58]. However, to the best of our knowledge, there are no similar knowledge-based approaches for measuring semantic relatedness between Arabic terms. Therefore, we selected a benchmark dataset that has been used in works on English text, and translated the terms included in the dataset to Arabic. The benchmark dataset includes human judgements on the similarity between the given terms. Thus, we use human-assigned judgments as a baseline to assess the accuracy of our approach. We also compared our approach with previous approaches that used the same dataset in English.

A. Benchmark Dataset

The used dataset is one of the WordSimilarity-353 test collection [59] that contains two sets of English term pairs along with human-assigned relatedness judgments. We selected the first set (set 1) that contains 153 term pairs along with their relatedness scores from 13 human subjects. The relatedness is assessed by human subjects by using a scale that ranges from 0 to 10 where 10 indicates the highest relatedness. Set 1 of the WordSimilarity-353 dataset also includes the list of 30 noun pairs from Miller and Charles [60]. To use it in our work, we translated the terms in set 1 to Arabic. The translation was carried out by the authors and was reviewed by an expert translator. Table 2 shows sample terms of WordSimilarity-353 after being translated into Arabic, along with the assigned human judgment scores. as an instance from the translated dataset.

TABLE 2
Snapshot of the translated WordSimilarity-353 dataset

Term 1	Term 2	Human Judgment													
		1	2	3	4	5	6	7	8	9	10	11	12	13	mean
نمر	قط	9	7	8	7	8	9	8.5	5	6	9	7	5	7	7.35
قطار	سيارة	7	7.5	7.5	5	3	6	7	6	6	6	9	4	8	6.31
الأوراق المالية	اليغور	1	0	0	1	4	0	2	0	1	3	0	0	0	0.92
الفيزياء	بروتون	9	8.5	9	10	6	8	8.5	8	7	8	9.5	5	9	8.12
مال	بنك	9	8	9.5	9	6	9	8.5	9	8.5	10	8	7	9	8.5
ساحل	هضبة	6	6	6	5	2	6	5	5	4	3	4	1	4	4.38
كوب	قهوة	9	8	9	8	5	9	8	5	6.5	5	4	3	6	6.58

We checked the existence of Wikipedia articles corresponding to the translated terms. Term pairs that do not have corresponding articles were excluded from the dataset because our approach computes semantic relatedness based on the presence of Wikipedia articles, i.e., each term can be mapped to a Wikipedia article. 23 out of the 153 pairs in the dataset were excluded, ending with 120 pairs. In addition, the mean value of human judgement scores was normalized to be in the range from 0 to 1 so that it becomes comparable with the normalized scores from our

approach. The complete translated dataset can be downloaded from: https://github.com/BaselAlhaj/SemanticRelatedness.

B. Experimental Conditions

Recall that our approach for measuring the semantic relatedness between terms uses two types of relations: the context relation and the category relation. For the computation of the category relation, two methods are used to compute the category's information value, which are: the depth-based information value and the descendants-based information value. Given that, our aim is to assess five different variants of our approach in order to find which setting gives most accurate results. These variants are as follows:

- **Sem-Context:** In this version, the semantic relatedness is measured by using only the context-based relation.
- **Sem-Category-Depth:** In this version, the semantic relatedness is measured by using only the category relation, where the category's information value is computed based on the category depth.
- **Sem- Category-Desc:** In this version, the semantic relatedness is measured by using only the category relation, where the category's information value is computed based on the category's descendants.
- **Sem-Context-Category-Depth:** This variant combines both context and category relations, where the category's information value is calculated based on the depth of the category.
- **Sem-Context-Category-Desc:** This variant combines both context and category relations, where the category's information value is calculated based on the descendants of the category.

C. Evaluation Metric

Results were evaluated by measuring the Pearson correlation [61] between the relatedness scores of our approach and the human judgement scores based on the following equation:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (7)$$

Where n is sample size, x_i, y_i are the individual sample points indexed with i and $\bar{x}, \bar{y}$ is the mean of x, y values respectively in the sample. The value of r is located in the range between +1 and −1, where $r = 1$ means a total positive correlation, $r = 0$ means no correlation exist, and $r = -1$ means a total negative correlation.

D. Results and Discussion

Table 3 shows the results from the five variants of our approach, in terms of Pearson correlation with human-assigned scores. In general, the version named Sem-Context-Category-Depth outperforms all over variants with a correlation of 0.68. The difference between Sem-Context-Category-Depth and other variants is statistically significant with $p < 0.05$ based on pairwise t-test. This indicates that the best setting for our approach is to combine both context and category relations, and to calculate the category's information value based on the depth of the category. This result also indicates that combining both the context relation and the category relation in our approach gives better results than using either of the two relations separately. When relations are used separately, we find that the Sem-Context version surpasses both the Sem-Category-Depth and the Sem-Category-Desc, but the difference was statistically insignificant. This indicates that when relations are used separately, the context-based relation is slightly more effective than the category-based relation.

TABLE 3
Experimental results in terms of the correlation with the human judgments

Variant of our approach	Correlation
Sem-Context	0.62
Sem-Category-Depth	0.60
Sem- Category-Desc	0.58
Sem-Context-Category-Depth	**0.68**
TABLE 3Sem-Context-Category-Desc	0.64

E. Comparison with Existing Approaches

One objective of the evaluation is to explore how our approach compares to other popular approaches from the literature that used the same dataset but with the English version of Wikipedia as a background knowledge. We compare our work with the following works that were discussed in the related works section: WikiRelate [57], WANG [62], SSA [63], WikiSim [64] and CPRel [65], ESA [58], WLM [3], WLVM [66] and WLA [67]. Table 4 shows the correlation values for all approaches. Figure 7 depicts the correlation values of compared approaches. Looking at the results, we notice that our approach performs better than some previous approaches such as: WikiRelate [57], WANG [62], SSA [63], WikiSim [64], and CPRel [65]. In contrast, it does not perform as well as other approaches such as: ESA [58], WLM [3], WLVM [66] and WLA [67].

We furtherly analyzed errors to better understand the reason behind these differences. In total, more than 70% of the reported errors occurred due to the poor content of the Arabic Wikipedia articles and the lack of links between the articles compared to the English version of Wikipedia. This lack of links hindered the computation of the context-based relation which primarily depends on the shared incoming links between articles. For example, the relatedness score obtained for the terms "مال" and "بنك" is 0.2, which is obviously inaccurate. The inspection of this error showed that the shared incoming links between the articles corresponding to these terms was only 135, comparing to 3623 links in the corresponding English articles. In addition, a lot of errors originated from the low complexity of the category graph in the Arabic Wikipedia compared to

the category graph of the English Wikipedia. The category graph is essential in our approach to estimate the information value of Wikipedia categories and to compute the category-based relation between terms. We found that the calculated information values for several categories differ across the two versions of Wikipedia and were mostly lower in the Arabic Wikipedia. In fact, English Wikipedia has about 3 times the number of categories and 2.8 times the number links between categories when compared to the Arabic Wikipedia[68]. To conclude, we believe that the difference in performance between our approach and others that rely on English Wikipedia can be mainly attributed to the gap between the Arabic and English versions of Wikipedia in terms of information richness and complexity.

TABLE 4
The proposed approach compared to other approaches.

Approach	Correlation
The proposed approach	0.68
Wikipedia Links and Abstract (WLA)[67]	0.72
Context Profile based Relatedness (CPRel)[65]	0.53
WikiSim [64]	0.63
Semantic Relatedness between Words based on Wikipedia Links [62]	0.63
Salient Semantic Analysis (SSA) [63]	0.622
Wikipedia Link-based Measure (WLM) [3]	0.69
Explicit Semantic Analysis (ESA) [58]	0.75
Wikipedia Link Vector Model (WLVM) [66]	0.72
WikiRelate [57]	0.49

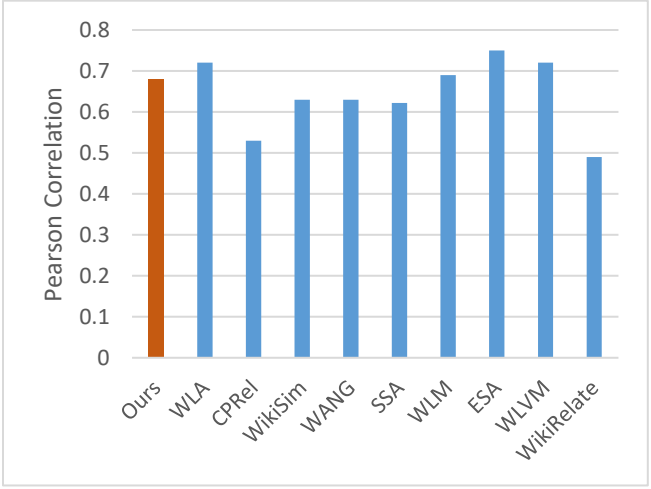

Figure 7. Comparison between the proposed approach and other similarity approaches that used the English version of Wikipedia

Conclusion and Future Work

In this work we proposed an approach for measuring the semantic relatedness between Arabic terms by exploiting Arabic Wikipedia as a knowledge source. Given two Arabic terms the approach selects the corresponding Wikipedia articles and uses their incoming links and categories to estimate the relatedness between them. Our approach was evaluated using a dataset from WordSimilarity-353 test collection which contains 120 pairs of terms with their human-assigned judgment scores. The results of the approach were compared with the results of human judgment and the results of other approaches that used the English version of Wikipedia. The correlation between our results and the result of human judgment was 0.68, which outperformed the results of some previous approaches that used the same dataset with English Wikipedia. The investigation of results has shown that many errors resulted from the lack of content of many Wikipedia articles and the poor category structure. This indicates that the Arabic version of Wikipedia can give satisfactory results when used as a background knowledge for semantic similarity measures, but it is still not as reliable as the English version.

References

[1] A. Budanitsky and G. Hirst, "Evaluating wordnet-based measures of lexical semantic relatedness," *Computational Linguistics,* vol. 32, no. 1, pp. 13-47, 2006.

[2] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," *International Journal of Computer Applications,* vol. 68, no. 13, pp. 13-18, 2013.

[3] I. H. Witten and D. N. Milne, "An effective, low-cost measure of semantic relatedness obtained from Wikipedia links," in *Conference of Association for the Advancement of Artificial Intelligence (AAAI)*, Chicago, the US, 2008, pp. 25-30.

[4] M. J. Hussain, S. H. Wasti, G. Huang, L. Wei, Y. Jiang, and Y. Tang, "An approach for measuring semantic similarity between Wikipedia concepts using multiple inheritances," *Information Processing & Management,* vol. 57, no. 3, pp. 102-188, 2020.

[5] Y.-w. ZHANG, B.-a. LI, X.-q. LV, S. Ning, T. Jing-Jing, and Engineering, "Research on domain term dictionary construction based on Chinese Wikipedia," in *International Conference on Applied Mechanics, Mathematics, Modeling and Simulation (AMMMS 2018)*, Hong Kong, 2018, pp. 225-230.

[6] P. Arnold and E. Rahm, "Extracting semantic concept relations from wikipedia," in *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)*, 2014, pp. 26-37: ACM.

[7] P. Arnold and E. Rahm, "Automatic extraction of semantic relations from wikipedia," *International Journal on Artificial Intelligence Tools,* vol. 24, no. 2, pp. 1-36, 2015.

[8] J.-X. Huang, K. S. Lee, K.-S. Choi, and Y.-K. Kim, "Extract Reliable Relations from Wikipedia Texts for Practical Ontology Construction," *Computación y Sistemas,* vol. 20, no. 3, pp. 467-476, 2016.

[9] Z. Wu *et al.*, "An efficient Wikipedia semantic matching approach to text document classification," *Information Sciences,* vol. 393, pp. 15-28, 2017.

[10] D. Chandrasekaran and V. Mago, "Evolution of Semantic Similarity—A Survey," *ACM Computing Surveys (CSUR),* vol. 54, no. 2, pp. 1-37, 2021.

[11] P. Sunilkumar and A. P. Shaji, "A Survey on Semantic Similarity," in *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*, 2019, pp. 1-8: IEEE.

[12] D. Sánchez, M. Batet, D. Isern, and A. Valls, "Ontology-based semantic similarity: A new feature-based approach," *Expert systems with applications,* vol. 39, no. 9, pp. 7718-7728, 2012.

[13] R. Navigli and S. P. Ponzetto, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artificial intelligence,* vol. 193, pp. 217-250, 2012.
[14] S. Banerjee and T. Pedersen, "Extended gloss overlaps as a measure of semantic relatedness," in *International Joint Conference on Artificial Intelligence*, 2003, vol. 3, pp. 805-810: Citeseer.
[15] Y. Jiang, X. Zhang, Y. Tang, and R. Nie, "Feature-based approaches to semantic similarity assessment of concepts using Wikipedia," *Information Processing & Management,* vol. 51, no. 3, pp. 215-234, 2015.
[16] G. Zhu and C. Iglesias, "Computing semantic similarity of concepts in knowledge graphs," *IEEE Transactions on knowledge and data engineering,* vol. 29, no. 1, pp. 72-85, 2016.
[17] D. Sánchez, M. Batet, and D. Isern, "Ontology-based information content computation," *Knowledge-based systems,* vol. 24, no. 2, pp. 297-303, 2011.
[18] M. A. Rodriguez, M. Egenhofer, and d. engineering, "Determining semantic similarity among entity classes from different ontologies," *IEEE transactions on knowledge,* vol. 15, no. 2, pp. 442-456, 2003.
[19] Y. Jiang, W. Bai, X. Zhang, and J. Hu, "Wikipedia-based information content and semantic similarity computation," *Information Processing & Management,* vol. 53, no. 1, pp. 248-265, 2017.
[20] J.-B. Gao, B.-W. Zhang, and X.-H. Chen, "A WordNet-based semantic similarity measurement combining edge-counting and information content theory," *Engineering Applications of Artificial Intelligence,* vol. 39, pp. 80-88, 2015.
[21] S. T. Dumais and Technology, "Latent semantic analysis," *Annual Review of Information Science,* vol. 38, pp. 189-230, 2004.
[22] S. Jain, K. Seeja, and R. Jindal, "Computing semantic relatedness using latent semantic analysis and fuzzy formal concept analysis," *International Journal of Reasoning-based Intelligent Systems,* vol. 13, no. 2, pp. 92-100, 2021.
[23] I. AlAgha and Y. Abu-Samra, "Tag Recommendation for Short Arabic Text by Using Latent Semantic Analysis of Wikipedia," *Jordanian Journal of Computers and Information Technology,* vol. 6, no. 02, pp. 165-181, 2020.
[24] A. Rozeva and S. Zerkova, "Assessing semantic similarity of texts–methods and algorithms," in *AIP Conference Proceedings*, 2017, vol. 1910, no. 1, p. 060012: AIP Publishing LLC.
[25] P. Mandera, E. Keuleers, and M. Brysbaert, "How useful are corpus-based methods for extrapolating psycholinguistic variables?," *Quarterly Journal of Experimental Psychology,* vol. 68, no. 8, pp. 1623-1642, 2015.
[26] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation," *arXiv preprint arXiv:.00055,* 2017.
[27] T. Kajiwara and M. Komachi, "Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 1147-1158.
[28] H. Jelodar *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications,* vol. 78, no. 11, pp. 15169-15211, 2019.
[29] R. Ben Djemaa, H. Nabli, I. Amous Ben Amor, and Experience, "Enhanced semantic similarity measure based on two-level retrieval model," *Concurrency and Computation: Practice,* vol. 31, no. 15, p. e5135, 2019.
[30] O. Araque, G. Zhu, and C. A. Iglesias, "A semantic similarity-based perspective of affect lexicons for sentiment analysis," *Knowledge-Based Systems,* vol. 165, pp. 346-359, 2019.
[31] H. T. Nguyen, P. H. Duong, and E. Cambria, "Learning short-text semantic similarity with word embeddings and external knowledge sources," *Knowledge-Based Systems,* vol. 182, p. 104842, 2019.
[32] I. Lopez-Gazpio, M. Maritxalar, M. Lapata, and E. Agirre, "Word n-gram attention models for sentence similarity and inference," *Expert Systems with Applications,* vol. 132, pp. 1-11, 2019.
[33] T. Zheng *et al.*, "Detection of medical text semantic similarity based on convolutional neural network," *BMC medical informatics and decision making,* vol. 19, no. 1, pp. 1-11, 2019.
[34] E. L. Pontes, S. Huet, A. C. Linhares, and J.-M. Torres-Moreno, "Predicting the semantic textual similarity with siamese CNN and LSTM," *arXiv preprint arXiv:.10641,* 2018.
[35] L. Yao, Z. Pan, and H. Ning, "Unlabeled short text similarity with LSTM encoder," *IEEE Access,* vol. 7, pp. 3430-3437, 2018.
[36] S. Zhang, X. Xu, Y. Tao, X. Wang, Q. Wang, and F. Chen, "Text Similarity Measurement Method Based on BiLSTM-SECapsNet Model," in *2021 6th International Conference on Image, Vision and Computing (ICIVC)*, 2021, pp. 414-419: IEEE.
[37] J. Kleenankandy and K. A. Nazeer, "Recognizing semantic relation in sentence pairs using Tree-RNNs and Typed dependencies," in *2020 6th IEEE Congress on Information Science and Technology (CiSt)*, 2021, pp. 372-377: IEEE.
[38] Y. Zhang, R. Tang, and J. Lin, "Explicit pairwise word interaction modeling improves pretrained transformers for english semantic similarity tasks," *arXiv preprint arXiv:.02847,* 2019.
[39] M. T. Elhadi, "Arabic News Articles Classification Using Vectorized-Cosine Based on Seed Documents," *Journal of Advances in Computer Engineering Technology,* vol. 5, no. 2, pp. 117-128, 2019.
[40] M. Belazzoug, M. Touahria, F. Nouioua, and M. Brahimi, "An improved sine cosine algorithm to select features for text categorization," *Journal of King Saud University-Computer and Information Sciences,* vol. 32, no. 4, pp. 454-464, 2020.
[41] M. O. Alhawarat, H. Abdeljaber, and A. Hilal, "Effect of stemming on text similarity for Arabic language at sentence level," *PeerJ Computer Science,* vol. 7, pp. 530-540, 2021.
[42] D. Schwab, "Semantic similarity of arabic sentences with word embeddings," in *Third arabic natural language processing workshop*, 2017, pp. 18-24.
[43] M. Bekkali, A. Lachkar, and Mining, "An effective short text conceptualization based on new short text similarity," *Social Network Analysis,* vol. 9, no. 1, pp. 1-11, 2019.
[44] F. A. Almarsoomi, J. D. OShea, Z. Bandar, and K. Crockett, "AWSS: An algorithm for measuring Arabic word semantic similarity," in *2013 IEEE international conference on systems, man, and cybernetics*, 2013, pp. 504-509: IEEE.
[45] H. Froud, A. Lachkar, and S. A. Ouatik, "Stemming versus Light Stemming for measuring the simitilarity between Arabic Words with Latent Semantic Analysis model," in *2012 Colloquium in Information Science and Technology*, 2012, pp. 69-73: IEEE.
[46] Y. Li, Z. A. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Transactions on knowledge and data engineering,* vol. 15, no. 4, pp. 871-882, 2003.
[47] G. Zakria, M. Farouk, K. Fathy, and M. N. Makar, "Relation extraction from arabic wikipedia," *Indian Journal of Science Technology,* vol. 12, pp. 46-52, 2019.
[48] A. M. Al-Zoghby, A. Elshiwi, and A. Atwan, "Semantic relations extraction and ontology learning from Arabic texts—a survey," in *Intelligent Natural Language Processing: Trends and Applications*: Springer, 2018, pp. 199-225.
[49] F. B. Mesmia, K. Haddar, D. Maurel, and N. Friburger, "Arabic named entity recognition process using transducer cascade and Arabic Wikipedia," in *Proceedings of the International Conference Recent Advances in Natural Language Processing*, 2015, pp. 48-54.
[50] M. Biltawi, A. Awajan, S. Tedmori, and A. Al-Kouz, "Exploiting multilingual wikipedia to improve arabic named entity resources," *International Arab Journal on Information Technology,* vol. 14, no. 4A, pp. 598-607, 2017.
[51] A. Alahmadi, A. Joorabchi, and A. E. Mahdi, "Combining Words and Concepts for Automatic Arabic Text Classification," in *International Conference on Arabic Language Processing*, 2017, pp. 105-119: Springer.

[52] A.-A. Iyad and A. Ahmed, "Towards Supporting Exploratory Search over the Arabic Web Content: The Case of ArabXplore," *Journal of Information Technology Management,* vol. 12, no. 4, pp. 160-179, 2020.
[53] Y. Ni *et al.*, "Semantic documents relatedness using concept graph representation," in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 2016, pp. 635-644: ACM.
[54] I. Traverso, M.-E. Vidal, B. Kämpgen, and Y. Sure-Vetter, "GADES: a graph-based semantic similarity measure," in *Proceedings of the 12th International Conference on Semantic Systems*, 2016, pp. 101-104: ACM.
[55] B. Louie, S. Bergen, R. Higdon, and E. Kolker, "Quantifying protein function specificity in the gene ontology," *Standards in genomic sciences,* vol. 2, no. 2, p. 238, 2010.
[56] T. Zesch, C. Müller, and I. Gurevych, "Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary," in *LREC*, 2008, vol. 8, no. 2008, pp. 1646-1652.
[57] M. Strube and S. P. Ponzetto, "WikiRelate! Computing semantic relatedness using Wikipedia," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2006, vol. 6, pp. 1419-1424.
[58] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *The International Joint Conference on Artificial Intelligence*, 2007, vol. 7, pp. 1606-1611.
[59] L. Finkelstein *et al.*, "Placing search in context: The concept revisited," *ACM Transactions on information systems,* vol. 20, no. 1, pp. 116-131, 2002.
[60] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity," *Language and cognitive processes,* vol. 6, no. 1, pp. 1-28, 1991.
[61] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*: Springer, 2009, pp. 1-4.
[62] R.-Q. WANG "Measuring of Semantic Relatedness between Words based on Wikipedia Links," *International Proceedings of Computer Science & Information Technology,* vol. 50, 2012.
[63] S. Hassan and R. Mihalcea, "Semantic Relatedness Using Salient Semantic Analysis," in *The Twenty-Fifth Conference on Artificial Intelligence (AAAI-11)*, 2011, pp. 884-889: San Francisco, CA.
[64] S. Jabeen, X. Gao, and P. Andreae, "Harnessing wikipedia semantics for computing contextual relatedness," in *Pacific Rim International Conference on Artificial Intelligence*, 2012, pp. 861-865: Springer.
[65] S. Jabeen, X. Gao, and P. Andreae, "CPRel: Semantic Relatedness Computation Using Wikipedia based Context Profiles," *Research in Computing Science,* vol. 70, pp. 57-68, 2013.
[66] D. Milne, "Computing semantic relatedness using wikipedia link structure," in *Proceedings of the new zealand computer science research student conference*, 2007, pp. 63-70.
[67] D. Zhao, L. Qin, P. Liu, Z. Ma, and Y. Li, "Computing terms semantic relatedness by knowledge in Wikipedia," in *Web Information System and Application Conference (WISA), 2015 12th*, 2015, pp. 107-111: IEEE.
[68] W. Lewoniewski, K. Węcel, and W. Abramowicz, "Multilingual ranking of wikipedia articles with quality and popularity assessment in different topics.," *Computers,* vol. 8, no. 3, pp. 1-32, 2019.