I.J. Information Engineering and Electronic Business, 2016, 1, 37-47
Published Online January 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2016.01.05

Exploring Semantic Relatedness in Arabic Corpora using Paradigmatic and Syntagmatic Models

Adil Toumouh
Computer Science department, Djillali Liabes University, Sidi Bel Abbes, 22000, Algeria
E-mail: toumouh@gmail.com

Dominic Widdows
Microsoft Bing, Bellevue WA, 98004, USA
E-mail: dwiddows@gmail.com

Ahmed Lehireche
Computer Science department, Djillali Liabes University, Sidi Bel Abbes, 22000, Algeria
E-mail: elhir@univ-sba.dz

Abstract—In this paper we explore two paradigms: firstly, paradigmatic representation via the native HAL model including a model enriched by adding word order information using the permutation technique of Sahlgren and al [21], and secondly the syntagmatic representation via a words-by-documents model constructed using the Random Indexing method. We demonstrate that these kinds of word space models which were initially dedicated to extract similarity can also been efficient for extracting relatedness from Arabic corpora. For a given word the proposed models search the related words to it. A result is qualified as a failure when the number of related words given by a model is less than or equal to 4, otherwise it is considered as a success. To decide if a word is related to other one, we get help from an expert of the economic domain and use a glossary[1] of the domain. First we begin by a comparison between a native HAL model and term- document model. The simple HAL model records a better result with a success rate of 72.92%. In a second stage, we want to boost the HAL model results by adding word order information via the permutation technique of sahlgren and al [21]. The success rate of the enriched HAL model attempt 79.2 %.

Index Terms—Relatedness, syntagmatic model, paradigmatic model, HAL model, term document model, word order information, permutation, Arabic corpus.

I. Introduction

Many NLP applications require knowledge that goes beyond similarity [5]. Thus, semantic relatedness is defined to cover any kind of lexical or functional association that may exist between two words [5]. Semantic relatedness determines whether two words are in some way related, even if they are not similar or have different parts-of-speech. Consequently, semantic similarity is a special case of the broader defined semantic relatedness, i.e. two words that are similar are also related, but the inverse is not true [28].

Humans can easily judge the semantic relatedness between two words. For example, they can easily tell that car and drive are strongly related, while there is no such strong connection between car and eat. This human ability is backed by their experience and knowledge, which makes it a hard task for machines. If a machine should solve this task, it also needs some knowledge source [28]. Different kinds of background information are used: structured resources such as WordNet (or any taxonomic resource) and ontologies where concepts are linked via semantics or lexical relations, encyclopedic resources like Wiktionary and Wikipedia; and unstructured resources, corpora being the predominant element in this area research. This work is concerned with this kind or background information.

The most popular models adopted for extracting meaning similarities from large text data are word space models. Based on the distributional hypothesis of Zelig Harris, linear algebra models are constructed by collecting context vectors for each word. The vectors reflect information about co-occurrence of words, so the similarity between two words is calculated as a distance measure between their representing context vectors. In general these models were designed for extracting similarity in term of synonymy; our aim is to explore the ability of this kind of model to capture the semantic relatedness between words in corpora. As it was explained by Sahlgren in [20] and [21] two predominant semantic relations exist: paradigmatic relations and syntagmatic relations. The first one is produced by the HAL-type model which collects co-occurrence data in

[1] www.pbf.org.ps/site/files/files/ موسوعة%20المصطلحات%20الاقتصادية.pdf

words-by-words information, while the second are concretized by the methods based on words-by-documents matrix, as example we cite LSA [12] and random indexing [10]. These models were criticized as "bag of words" that encodes only the contexts in which words co-occur, but ignore word-order information [16]. The pioneering methods have adapted the HAL model (paradigmatic representation) to encode word-order information: the first one was proposed by Jones and Mewhort [7] based on convolution operation; and the second is based on permutation of random vectors and proposed by Sahlgren and al. [21]. Functionally, the two approaches are quite similar, but random permutation is much more computationally efficient than convolution [4].

Our work is concerning by exploring the two paradigms: paradigmatic representation via the HAL model enriched with word order information using the permutation technique of sahlgren and al [21], and the syntagmatic representation via a words-by-documents model constructed using the random Indexing method. In this work, we have chosen to practice our experiments on Arabic corpus. This choice is motived by our conviction that Arabic language is one of many less widespread languages which deserves to have more attention to get more benefits from these new methods, for example more efficient tools for constructing and maintaining lexical and semantic resources.

II. Geometric Models for Distributional Semantics

Semantic vector models have received considerable attention from researchers in natural language processing over the past 15 years, though their invention can be traced at least to Salton's introduction of the Vector Space Model for information retrieval [22][23][31]. The underlying assumption is motivated by the works of Zellig Harris and known as the *distributional hypothesis.* Harris's idea was that the words with similar distribution in language have similar meaning, and therefore can be grouped according to their distributional behavior [20][21]. The core idea behind semantic vector models is that words and concepts are represented by points in a mathematical space, and this representation is learned from text in such a way that concepts with similar or related meanings are near to one another in that space [31]. These semantic representations of words are achieved by collecting statistical redundancies observed in a large corpus of text [16]. Then the semantic similarity or relatedness is quantified by comparing their distributional profiles. Geometrically speaking, words and concepts are represented by vectors in high-dimensional space. This is normally done by collecting word occurrence frequencies in high-dimensional context vectors [21]. Therefore, the similarity between word meanings is expressed by quantifying the proximity between the vectors representing the words [31].

Gathering Co-occurrence Data

The most standard form of co-occurrence data in this field is the traditional term-document matrix used in information retrieval [32] [23]. For a large corpus of documents, a term-document matrix records the number of times each term occurred in each document. We can think of this either as describing each document as a collection of terms, so that the documents are the objects and the terms are the features. However, we can also think of this as describing a term as a collection of documents – in which case, the terms are the objects and the documents are the features.

The Hyperspace Analogue of Language (HAL) models [13] used in the Stanford Infomap project and the Infomap-NLP software [24][29], are an another form of co-occurrence data, in which the features used were other words in the text (often chosen to be words of high frequency occuring within a small window of words surrounding each target word), so that instead of capturing term-document counts, the matrix captures term-term counts. In general, the method therefore consists of collecting the distributional statistics of words into a co-occurrence matrix. Each row (respectively column) stands for a unique word (respectively context) and records its co-occurrences in the different contexts (respectively for the different words) (Fig.1). These frequencies are often normalized by using for example the TF-IDF weighting score. The matrix is referenced as a words-by-documents matrix when the contexts are documents and words-by-words matrix when the context is reduced to words [19]. The dimensionality of the WordSpace is determined by the number of columns. Each row is called a context vector. Computing similarity between words become as computing similarity between their vectors representations by using a kind of similarity measures, such as the cosine similarity.

W_i: words
Cj: contexts (word, document,…)
fij: frequency of co-occurrence

	C_1	C_2	C_3
W_1	12	00	7
W_2	22	02	03
W_3	01	09	18
W_4	00	05	06

Fig.1. Example of Co-Occurrence Matrix.

Unfortunately, the number of the documents we manipulate are very large, which make the dimensionality of the matrix very high. Consequently, the matrix will be intractable and affects the scalability of this method. The solution is to preserve the use of large collection of text but represented in low dimensional context vector. This is the subject of sections III and IV.

III. Random Indexing, as a Base Model and Matrix Reduction Tool

A practical problem with the co-occurrence matrix is that as the number of documents increases, the

dimensionality of the space grows. With a high-dimensional space, this solution can become computationally intractable and very sparse. An ideal solution is to preserve a big size of data and to project them in lower dimensional space. However, reduction of dimensionality will affect the integrity of document information. For this reason, research has been undertaken focused on reducing the dimensionality efficiently while preserving as much information from documents. These dimension reduction techniques are applied as a data pre-processing step. Several techniques have been developed, including 'latent semantic analysis' (LSA) [12] and 'Hyperspace Analog to Language' (HAL) [13]. Unfortunately, these methods are based on matrix factorization techniques, which are costly operations, and they cannot be applied before constructing the huge matrix, which makes it difficult to support incremental addition of new terms and documents to models without rebuilding them from scratch. So, such methods remain computationally heavy, and thus are liable to efficiency and scalability issues. As an alternative to such computationally heavy dimensionality-reduction techniques, many approaches based on the Johnson-Lindenstrauss lemma [6] have been developed:

For any $0< \varepsilon <1$ and any integer n, let k be a positive integer such that $k \geq 4(\varepsilon^2 / 2 - \varepsilon^3 / 3)^{-k} \ln n$

Then for any set W of n points in R^d, there is a map f: $R^d \rightarrow R^k$ such that for all $u, v \in W$,

$$(1- \varepsilon)||u-v||^2 \leq ||f(u)-f(v)||^2 \leq (1+ \varepsilon)||u-v||^2$$

The Johnson-Lindenstrauss lemma says: if we project points of a vector space into a randomly selected subspace with sufficiently high dimensionality, the distances between the points are approximately preserved. This means that we can approximate the orthogonality by simply choosing random directions in the high-dimensional space. This near-orthogonality of the random vectors is the key to a family of dimension reduction techniques that includes methods such as random projection [14], random mapping [11] and random indexing [10]. The random projection methodology process is as follows: each context is assigned a unique and random index vector of ternary values (0, +1, –1). These vectors consist of a large number of 0s and a small number (about 1%–2%) of +1s and –1s randomly distributed. The dimensionality of these randomly generated vectors is usually chosen to be on the order of hundreds or thousands. As the corpus is scanned, the context vector for each word is obtained by summing the index vectors of all the contexts in which the word appears. For instance: suppose that we have three documents D1, D2, D3, each one is assigned respectively the three random index vectors: {00 + 100…} {–10000}{0 + 100 – 1…}. If a word w occurs in the three documents, the context vector of this word will be the sum of the three vectors: {–1 + 1 + 10 – 1…}. Random indexing has many advantages compared to others methods [17]:

- It is an incremental method, which means that we do not have to sample all the data before we can start using the context vectors
- It avoids the 'huge matrix step', since the dimensionality k (k << d) of the vectors is much smaller (d is the number of the columns in the original matrix)
- It is scalable, since adding new contexts to the data set does not increase the dimensionality of the context vectors.

IV. Encoding Word-Order Information

These kinds of semantic spaces are qualified as bag-of-words models: they succeed in encoding contextual information and fail in capturing information about how words are ordered in the sentence. In recent years, two approaches in building semantic spaces have been developed in order to take into account information about word-order: the Bound Encoding of the Aggregate Language Environment (BEAGLE) model [7] and a permutation model [21] based on Random Indexing [10].

As it was described in the section of Random Indexing, these two methods build the semantic spaces by using vectors generated randomly with fixed length (called environmental Vectors in BEAGLE model) intended to represent the invariant properties of each word, as well as dynamic memory vectors that store information about each word's semantic representation [16]. Each time a word is encountered in the corpus, its memory vector is updated with context information provided through the superposition of the environmental vectors for every other word in the surrounding sentence [4]. To Represent order information about the word, BEAGLE and RPM bind together collections of environmental Vectors into order vectors that are added to memory vectors during training [16]. The main challenge facing efforts to encode syntactic information into high-dimensional spaces is to find an appropriate, order-preserving mathematical operation for recursively combining vectors [16]. In other words, in a composite vector (vector containing contextual and order information), one must be able to determine which features (elements) go with which objects (original vectors)—this is referred to as the binding problem. Furthermore, we need also to determine the original ordering of the vectors from a composite representation [7]. The two proposed techniques address the problems via two different methods; Jones and Mewhort [7] employ a non-commutative circular convolution proposed by Plate [15], while the Sahlgren and al. technique is based on random permutations of the environmental vectors. These two techniques are explained in more detail as follows.

BEAGLE (Bound Encoding of the Aggregate Language Environment)

The Jones and Mewhort [7] approach is based on a convolution operation, at the end of the process, each word is assigned a single composite vector representation, a combination of context and order information [7]. At the beginning, a so called environmental vector e_i is

created for each term. Its components are sampled randomly from a Gaussian distribution with $\mu = 0$ and $\sigma = 1/\sqrt{D}$. In order to compute context and order information for each word *w*, we need three memory vectors:

c_i: Vector representing contextual information
o_i: Vector representing order information
m_i: Is the combination of c_i and o_i.

Each time a word is encountered, its environmental vector is used in the coding of its new context and order information, and this new information is added to m_i. $m_i = m_i + c_i + o_i$

Referring to Jones and Mewhort [7], the idea is: context information comes from vector addition, and order information comes from vector convolution. Therefore, context information is obtained by superposition (summation) of environmental vectors, while the order is sampled by forming association between words with vector convolution.

Permutation method of Sahlgren

The version of sahlgren and al [21] in encoding word-order information is based on random indexing technique. The environmental vectors are called Random index vectors, a high dimensional, random, sparse and ternary vector. The context information can be encoded as the same way in Jones and Mewhort approach [21]. Sahlgren and al use permutation or shuffling of coordinates (shifting of all of the non-zero values of a sparse elemental vector to the left or right according to the relative position of terms) to replace the convolution operator [8].

V. Experimentations and Analysis

The experiments we realized are subdivided into 2 sections:

1. The first one is concerned with a comparison between the term-by-document model (syntagmatic model) and the HAL model (paradigmatic model).
2. The second focus on the model HAL and the experimentation of introducing order information in the HAL model. We hope to discover if this new information contributes in improving the accuracies of the HAL model obtained in the first section of experiments. We note that all the models are built by using the technique of random indexing (section III).

These experiments are based on two main materials:

- The Khaleej-2004[2] (4.1 MB): A corpus extracted from the daily Arabic newspaper Akhbar Al Khaleej', it includes 5120 news articles corresponding to 2.855.069 words covering four topics: sport, local news, international news and economy. We opted for the economy topic which consists of 909 files [1].
- The semantic vectors[3]: an open source package characterized by its simplicity, ease of use, and scalability. The software can be used to easily create semantic vector models from a corpus of free text, and to search such models using a variety of mathematical operations including projections and algebraic product operations [31]. Due to its adaptability, many NLP applications and experiments have used it as a component. This software allows a range of possible applications: the most immediate perhaps is in measuring word similarity; semantic vector models can also be used in resource building applications such as ontology learning and lexical acquisition, as part of data gathering for decision-support systems, detailed research, etc. [29]

To perform the experiments, we selected 48 specific economic terms from: A GLOSSARY OF COMMON TERMS USED IN CUSTOMS, TRADE, ACCOUNTING AND ECONOMICS[4]:

5.1 HAL model VS Term-document model : Part I

This first part of the experiments consists in comparing two models: the HAL model and the term-document model. As it is demonstrated in the figure Fig. 2, the two models are built using the semantic vectors package and the core dimensionality reduction method is random indexing.

The HAL model needs a principal parameter which is the width of the moving window; it indicates the context to be taken in account in the left and the right for the target word. We have opted for varying this parameter from 1 to 7, furthermore that means we constructed 7 HAL model, each one corresponds to window radius having a value varying from 1 to 7.

Now, the 8 models: 7 HAL models and the term-document model are ready to be queried by the 48 selected words specific to customs, trade, accounting and economics (Fig 3).

Our aim is, for a given word we must extract the related economics words; this will be achieved by selecting the related words from the 20 words constituting the results set. Deciding if a word is related or not of the query word will be realized by an expert of economic domain and by consulting the GLOSSARY OF COMMON TERMS USED IN CUSTOMS, TRADE, ACCOUNTING AND ECONOMICS[5].

We have decided to consider the result wrong if it give us under 4 related terms for the query. For example, for the word قرض -it's translation in English is Loan-, the Hal model with a context window fixed to one word provides us only two related words: مشروع، برنامج -their translation are respectively: Project, program - While the same model with a moving window fixed to 4 provides us many close words: راسمال، اصدار، سند، انتاج، يورو، برنامج، مبلغ

[2] https://sites.google.com/site/mouradabbas9/corpora

[3] https://github.com/semanticvectors/semanticvectors

[4] www.pbf.org.ps/site/files/files/موسوعة%20المصطلحات%20الاقتصادية.pdf

[5] www.pbf.org.ps/site/files/files/موسوعة%20المصطلحات%20الاقتصادية.pdf

مشروع، اصول، مستثمر -their translation respectively are: Amount, Capital, Issue, Support, Production, Euro, Program, Project, Assets, Investor -.

For the term-by-document model the success is about 65%, while for the HAL model, the best score is 72.92% when the moving window is fixed to 3, and the worst one is 60, 42% when we fix the window radius to 1 and 7(table1).

مصرف (Bank)، سند (support)، إنفاق (spending)، دخل (income)، ربح (profit)، سعر(price)، مؤشر(index)، صفقة (deal)، تنمية (development)، فائدة (benefit)، برميل (Barrel)، نقد (cash)، مردودية (cost-effective)، خسارة (loss)، قرض (loan)، رهن (mortgage)، رصيد (credit)، كساد (recession)، ضرائب (taxes)، رسم (fee)، تكاليف (costs)، أصول (assets)، دولار(dollar)، البطالة (unemployment)، تصدير (export)، تهريب (smuggling)، بنك (bank)، أوراق (securities)، تجزئة (retail)، أوبك (OPEC)، وقود (fuel)، إتفاق (agreement)، إستثمار (investment)، فاتورة (bill) ، سوق (market)، مقايضة (swap)، سيولة (liquidity)، رأسمال (capital)، إدخار (savings)، تضخم (inflation)، بضاعة، (merchandise) إحتكار (monopoly)، إقتصاد (economy)، نفط (oil)، أسهم (Shares)، عجز (deficit)، ميزانية (budget)، معدل (Rate).

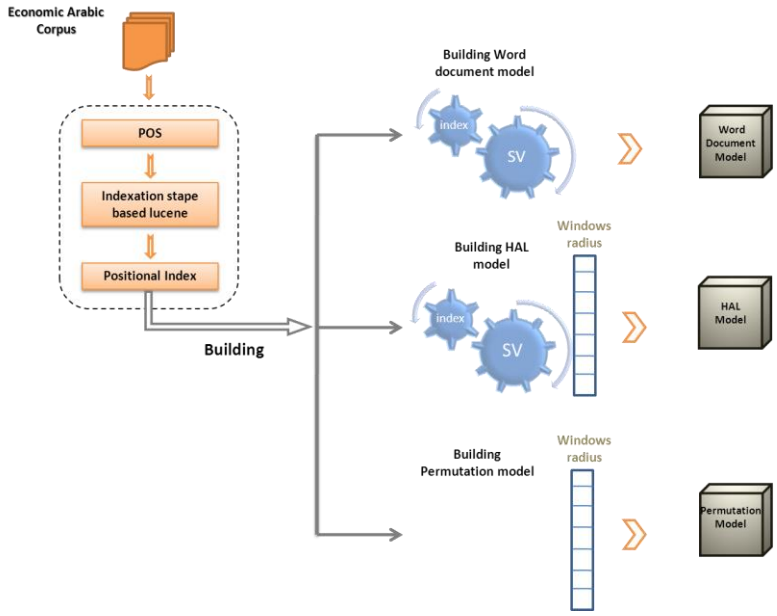

Fig.2. Constructing the Models of Experimentations

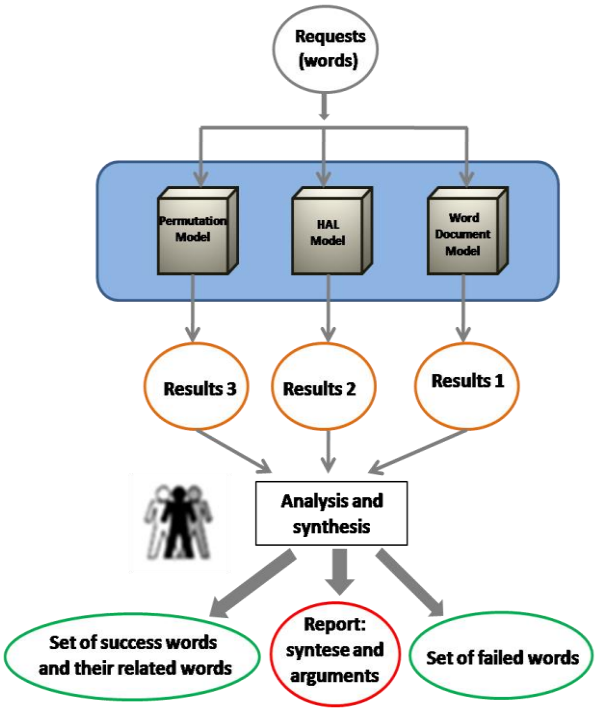

Fig.3. Requesting the Three Models

Table 1. Results of Term-By-Document and HAL Model

Term-Document model		success	31	64.59%
		failure	17	
HAL model	window radius =1	success	29	60.42%
		failure	19	
	window radius =2	success	34	70.33%
		failure	14	
	window radius =3	success	35	72.92%
		failure	13	
	window radius =4	success	34	70.33%
		failure	14	
	window radius =5	success	33	68.75%
		failure	15	
	window radius =6	success	34	70.33%
		failure	14	
	window radius =7	success	27	60.42%
		failure	21	

Table 2. Examples of Excellent Results for the Term-By-Document and HAL Models

Query term	Its translation in English	The model	Related words	Their translation
مصرف	bank	Term-document model	مال، مؤسسة، فروع، ائتمان، بنك، ضابط، اسلام، محافظ، دولار،مطالب، راسمال، املاك	Capital, Corporation, Branches, Credit, Bank, Officer, Islam, Governor, Dollars, Demands, Capital, Properties
		HAL model with window radius = 3	عقار، بنوك، اقتصاد، حكومة، سياح، مال، توسع، عمل، دولة، خدمة، مؤتمر، مؤسس، تطوير	Tourists, Capital, Drug, Banks, Econo my, Government, Expansion, Work, State, Service, Conference, Founder, Development
		HAL model with window radius = 6	مال، مؤسسة، استثمار، اسلام، خدمة، بنوك، عقار، شركة، مركز، عالم	Capital, Corporation, Investment, Islam, Service, Banks, Drug, Company, Center, World
سعر	price	Term-document model	دولار، طن، قنطار، شتاء، صويا، فول، برنت، نحاس، قمح، حاصل	Dollars, Tons, Quintals, Winter, Soy, Bean, Brent, Copper, Wheat, Holds
		HAL model with window radius = 3	، تراجع انتاج، يورو، شركة، اوبك، نسبة، ارتفاع، انخفاض، نفط، حكومة، راسمال، هبط	Produce, Euros, Company, OPEC, Percentage, Retreat, High, Low, Oil, Government, capital, Fall
		HAL model with window radius = 6	ارتفاع، تراجع، انخفاض، نفط، يورو، عقود، خام، نسبة، انتاج	High, Retreat, Low, Oil, Euros, Contracts, crude, Percentage, Produce
مؤشر	index	Term-document model	نفط، متداول، اسبوع، اقفال، تراجع، اصول، قراءة، اسهم،	Oil, Trader, Week, Locks, Retreat, Assets, Reading, Shares,
		HAL model with window radius = 3	،توقع، اقفل مستقبل، ،مملكة راسة ، مال، اداء، سجل سيسكو،	Capital, Study, Kingdom, Future, Closed, Expectation, Cisco, Record, Performance
		HAL model with window radius = 6	مستوى، نفط، اداء، سوق، تراجع، جديد، حالة،	Level, Oil, Performance, Market, Retreat, New, Status
تنمية	Development	Term-document model	قطاع، اقتصاد، مستقبل، شعوب، استثمار، دعم، فرص، اعمال، صناعة، سياسة	Sector, Economy, Future, Peoples, Investment, Support, Opportunities, Business, Industry, Policy
		HAL model with window radius = 3	دور، اجتماع نشاط، تطوير، تعزيز، سياسة، مستقبل، وطن، نمو ،دعم، استثمار	Development, Strengthen, Policy, Activity, Role, Meeting, Future, Investment, Support, Growth, Home
		HAL model with window radius =6	تطور، تكتل، تكامل، سياسة، تطور، اصلاح، انتعاش، نشاط، مردود، فلسفة، وطن	Development,In front of, Integration, Policy, Development, Reform, Recovery, Activity, Yields, Philosophy, Homeland

Analysis

1. We note that the HAL model gives better results than the term-document model when the size of window is varying from 2 to 6.
2. In many cases the results were excellent for the Hal model (with different moving windows) and the term-by- document model. The following examples shown in Table 2 demonstrate this remark.
3. Other important case to note, the related words given for the query word سعر (price in English) are very interesting, they are very specific to the economics domain and so close to the word سعر. The results set given by the term-document model is formed by many names of products: فول، صويا، برنت، شتاء، اسعار طن، قنطار نحاس، قمح، حاصل, their translations respectively in English are: winter, Prices, Tons, Quintals, Bean, Soy, Brent, Copper, Wheat, Holds. In the same way, we note many kind of interesting result; the term-document model gives us a set of related words specific to nouns of societies and banks.
4. While in other cases, the results were successful, but

the number of the related words were lower and they were not very close as the ones in previous table 2 (see Table 3).

5. We note that the best percentage were scored for the moving window varying from 2 to 6, and the best one was for the window radius=3.
6. The worst results were for the window radius =1 and 7 (60.42%), i.e.: When the context is limited and very small with a window radius =1 and when it is taken with a large set of words as consequence the results are altered.

5.2 HAL model based permutation : part II

The best result obtained in the experiments of Part I is the one of the Hal model with WR = 3. It is the only one which far exceeds the 70% of success (it recorded 72.92%). The purpose of this second part is to test the possibility of increasing the success rate beyond 75% and to reach the vicinity of 80% (which means that over three quarters are positive results).For this, we investigate the HAL model enriched with word-order information, it is known as the permutation based model (see section III) proposed by Sahlgren et al. [21].

Table 3. Examples of Lower Quality Results than Those of the Table 2

Query term	It's translation in English	The model	Related words	Their translation
معدل	Rate	Term-document model	نمو، ارتفاع، عجز، فائدة، تسارع، تذبذب، محلل، فيدرال، احتمال	Growth, High, Deficit, Usefulness, Acceleration, Fluctuation, Analyst, Federal, Prospect
		HAL model with window radius = 3	نسبة، نمو، تباطؤ، استمرار، تطور، اسعار، نشاط، شركة، انفاق، تراجع، زيادة	Percentage, Growth, Slowdown, Continuation, Evolution, Rates, Activity, Company, Spending, Retreat, Increase
		HAL model with window radius = 6	نسبة، نمو، زيادة، انخفاض، استمرار، حجم، اقتصاد، تطور	Percentage, Growth, Increase, Low, Continuation, Size, Economy, Development
فائدة	Interest	Term-document model	معدل، ايداع، تضخم، تنافس، كريدي، سنويا، مضاعفا	Rate, Deposit, Inflation, Compete, Credit, Annually, Doubly
		HAL model with window radius = 3	نفط، ارتفاع، انخفاض، تضخم، استمرار، انفاق، عائدات	Oil, High, Low, Inflation, Continuation, Spending, Revenues
		HAL model with window radius = 6	ارتفاع، انخفاض، نفط ، تضخم، طلب، عائدات، معدل، مستوى	High, Low, Oil, Inflation, Asked, Revenues, Rate, Level
ربح	Profit	Term-document model	بنك، بورصة، نمو، خسارة، عائدات، تسويق، اقرار	Bank, Stock Exchange, Growth, Loss, Revenues, Marketing, Approval
		HAL model with window radius = 3	شركة، مساهم، مستثمر، نسبة، راسمال، اسهم، زيادة، حكومة، انخفاض، استثمار	Company, Shareholder, Investor, Percentage, capital, Shares, Increase, Government, Low, Investment
		HAL model with window radius = 6	شركة، استثمار، مؤسسة، حكومة، عمل، مستثمر، منتج، نسبة، حجم	Company, Investment, Corporation, Government, Action, Investor, Product, Percentage, Size
رصيد	credit	Term-document model	تراكمت، عزوف، بنك، بيع، موزعا، مشترك، مزود	Accumulated, Reluctance, Bank, Selling, Distributor, Mutual, Provider
		HAL model with window radius = 3	زيادة، شركة، مستثمر، مساهم، حصة، مؤسسة، انجاز، مواطن	Increase, Company, Investor, Shareholder, Share, Corporation, Achievement, Citizen
		HAL model with window radius = 6	حكومة، مؤسسة، استراتيج، خبراء، مستثمر، نشاط، مشروع، نمو، زيادة، انفاق	Government, Corporation, Strategies, Experts, Investor, Activity, Project, Growth, Increase, Spending

Table 4. Results of Permutation Model (HAL Model Enriched with Order Information using the Permutation Method)

HAL model based permutation	window radius =1	success	25	52.1 %
		failure	23	
	window radius =2	success	35	72.92 %
		failure	13	
	window radius =3	success	34	70.83 %
		failure	14	
	window radius =4	success	35	72.92 %
		fail	13	
	window radius =5	success	36	75 %
		failure	12	
	window radius =6	success	38	79.2 %
		failure	10	
	window radius =7	success	36	75 %
		failure	12	

The results are very satisfactory, for WR = 6 the success rate reached almost 80% by recording 79.2%. For WR=5 and WR = 7 the success rate is 75%. For WR=2 and WR = 4, their success rate equalize the best score recorded in Part I (72.92 %).

Analysis

1. The most important remark and conclusion is that HAL model enriched with the information of word-order and based on the permutation method provides excellent results.

In most results considered positive, the provided related-words were very close and very connected to the requested word. We note also the high number or related-words compared with the results obtained in Part I.

2. The words for which a total failure was recorded regardless of the window radius are: بضاعة ، احتكار، مقايضة، سوق، فاتورة.
3. For the model based on permutation there are various cases where the results were excellent (in quality of results number) whatever to the moving window (from 1 to 7), we cite as example: اتفاق, سند, اصول, دولار, فائدة, تضخم. As an example we propose the word اتفاق and the different results obtained for all the sliding windows (ranging from 1 to 7) in Table5.

Table 5. Very Excellent Results of the word اتفاق (Deal) with All Window Radius

Query term and Its translation in English	Window radius	Related words	Their translation
اتفاق	1	مخاطر، صناديق، سيولة، خسائر مشاكل، مشاريع، تجارب، معوق، مزايا	Risks, Funds, Liquidity, Losses, Problems, Projects, Experiences, Blocker, Advantages
	2	شركة، مخاطر، سيولة، مسؤول، مشاريع، مكاتب، مستثمر، صناديق، منافسة، حكومة، عملاء، ايرادات، مساهم، خدمة	Company, Risk, Liquidity, Administrator, Investor, Boxes, Projects, Offices, Competition, Government, Clients, Revenue, Shareholder, Service
	3	مخاطر، محفظة، شركة، مستثمر، مشروع، اتفاق، عمل، مستقبل، طاقة، عائدات، برنامج	Risks, Purse, Company, Investor, Project, Agreement, Action, Future, Energy, Revenues, Program
	4	شركة، سند، استثمار، ايرادات، مشاريع، حجم، ارباح، انجازات	Company, Support, Investment, Revenue, Projects, Size, Profits, Achievements
	5	شركة، حكومة، صناديق، حجم، مستثمر، فرص، اسهم، تداولات	Company, Government, Boxes, Size, Investor, Opportunities, Shares, Trading
	6	شركة، حكومة، مستثمر، انتاج، مشاريع، انتاج، خدمة، استثمار، نسبة، منافسة، عملاء، عقارات	Company, Government, Investor, Produce, Projects, Produce, Service, Investment, Percentage, Competition, Clients, Property
	7	شركة، استثمار، حكومة، نسبة، مستثمر، مساهم، اسهم، ارباح، راسمال، مبادرة، تدفق، موظف، مخاطر	Company, Investment, Government, Percentage, Investor, Shareholder, Shares, Profits, Capital, Initiative, Flow, Employee, Risks

5.3 Synthesis of Part I and Part II

1. Before achieving the comparison between the three models and choosing which one success more in capturing the relatedness between words, our principal challenge was to explore and to discover if the semantic spaces are suitable for a such kind of task and specifically for many less widespread languages. The results obtained were very successful and prove the ability of these geometric and semantic space to capture the relatedness between words, the permutation based Hal model achieve 79.2% as a percentage of success, the simple Hal model realized about 73% of success, and the term-by-document model was satisfied with de 64,59 of success (Fig. 4). The most important point doesn't be resumed at the success obtained by these geometric models in capturing the similarity and the relatedness between words, but the fact of the reduced numbers of needed resources make these models very important and interesting. We have needed only a corpus and the simple distributional hypothesis for constructing our three models.
2. One very important remark is the agglutinated words given at many times by the term-by-document model. Some results sets contain words like: قطاعاتنا لطاقاتهم أنظمتهم مداخلاتهم محافظهم, their translation respectively in English are: our sectors, for their energies, their systems, interventions, portfolios. Words like قطاعاتنا أنظمتهم contain respectively the pronouns "نا " and "هم " (their translation respectively are : "our" and "their"), and the word like لطاقاتهم contain the preposition "لِ", these pronouns and prepositions constitute what we call the clitics, their adjunction to words product agglutinated words and bring us in the problem of clitization. In fact, we mustn't find this kind of words after a successful morphological analysis stage. That's means; the morphological analysis engine couldn't perform properly its task. The word like أنظمتهم (their systems) mustn't found in this form, the morphological analyzer must provides أنظمتwith the pronoun "هم"and not the complete agglutinated word أنظمتهم. This kind of misses' cases can alter the results of indexing stage and consequently the results of extracting related words.

3. The words of the used corpus does not include a set of special marks called diacritics, it is a known ¶problem encountered for the most of the Arabic manuscripts. ¶A great portion of the words of the Arab language accept a multitude of diacritics, for instance¶ the word سند -its translation English is deed- represent a many ambiguities when take in account the diacritics level (table 6). So, the words will have different meaning depending on how they are diactritized, which can induce a strong ambiguity and influences the semantic analysis such as the extraction of synonymy or the extraction of the relatedness in our case.

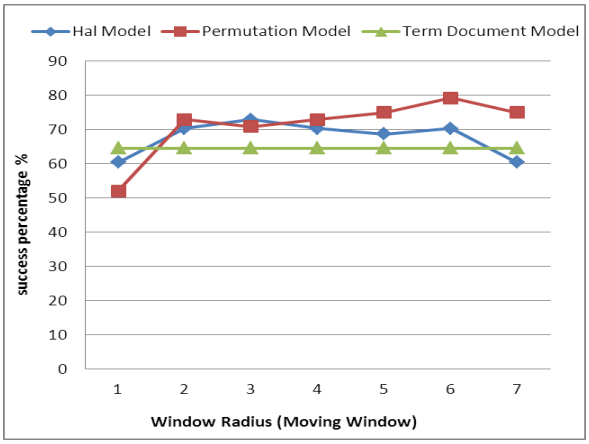

Fig.4. The Curves Representing the Results of the Three Models

Table 6. Examples of Diacritization Problem (Arabic Definitions Were Taken from the Electronic Dictionary Almaany[6])

The word and it's translation	Synonyms and Signification in Arabic	It's definition (signification) in English
سَنَّدَ (advocacy)	دَعَمَ . سنَده ، وثَّقه ، دعَّمه ، جعل له عِمادًا يرتكز عليه	lie or be placed against something for support
سَنَد (pillar)	دِعَامَة . كلّ ما يُعتمد عليه ويُستند إليه	bolster, crutch, hold, prop, rest, shore, stake, stay, support-- A long stout piece of timber or metal set upright in the ground to support something
سِند (Sindh)	السِّنْدُ : اسم مكان يطلق على الجزء الشمالي الغربي من الهند ، يتوسطه حوضُ نهر السّند ، وأكثرُه الآن يقع في باكستان	Place name given to the north-western part of India, strikethrough Indus River basin, and most now located in Pakistan
السِّنْدُ (Sindhis)	جِيلٌ يسكنُ تلك البلاد	People living in that country
السِّنْدُ (yemeni clothes)	ضربٌ من الثياب أَو البُرُودِ اليمانية	Kind of Yemeni clothes
السَّنَدُ (deed)	في الاقتصاد: ورقةٌ ماليَّةٌ مثبتةٌ لقرض حاصل ، وله فائدةٌ ثابتةٌ	In the economy: financial paper installed to loan holds, and has a fixed interest

[6] http://www.almaany.com/

VI. Conclusion

This work demonstrates the ability of Word Space Models based on random indexing for extracting the relatedness from Arabic corpora. We have explored three models, term-document model, HAL model and the permutation model (HAL model enriched with word order information). The best result was given by the last one with a success rate of 79, 2% when we fixed the moving window to 6. The developed approach is very simple and is not gourmand on resources, it needs only an Arabic corpora. The obtained results demonstrate the efficiency of using geometric models and specially the paradigmatic models for capturing the relatedness between words. This success can be exploited in process of constructing and enriching semantic resources like ontologies or linguistic resources for widespread languages like Arabic.

References

[1] M. Abbas and K. Smaili, "Comparison of Topic Identification Methods for Arabic Language," International conference RANLP05: Recent Advances in Natural Language Processing, 21-23 september 2005, Borovets, Bulgary.

[2] K. Ben Sidi Ahmed and A. Toumouh, "Effective Ontology Learning: Concepts' Hierarchy Building using Plain Text Wikipedia" ICWIT 2012: 170-178.

[3] K. Ben Sidi Ahmed, A. Toumouh and D. Widdows, "Lightweight domain ontology learning from texts: graph theory–based approach using Wikipedia," International Journal of Metadata, Semantics and Ontologies, Volume 9 Issue 2, April 2014, Pages 83-90.

[4] P. Blouw, and C. Eliasmith, "A Neurally Plausible Encoding of Word Order Information into a Semantic Vector Space," 35th Annual Conference of the Cognitive Science Society, 2013.

[5] A. Budanitsky and G. Hirst, "Evaluating WordNet-based Measures of Semantic Distance," Computational Linguistics, 32(1), 2006.

[6] W. B. Johnson, and d. J. Lindenstrauss, "Extensions of Lipshitz Mapping into Hilbert Space," In Conference in modern analysis and probability, volumn 26 of Contemporary Mathematics, pages 189-206. Amer. Math. Soc., (1984).

[7] M. N. Jones, and D. J. K. Mewhort, "Representing word meaning and order information in a composite holographic lexicon," Psychological Review, 114, 1-37 (2007).

[8] S. Jonnalagadda, T. Cohen, S. Tze-Inn Wu, and G. Gonzalez, "Enhancing clinical concept extraction with distributional semantics," Journal of Biomedical Informatics 45(1):129-140 (2012).

[9] P. Kanerva, "Sparse Distributed Memory", MIT Press, (1988).

[10] P. Kanerva, J. Kristofersson, and A. Holst, "Random Indexing of text samples for Latent Semantic Analysis," in Proceedings of the 22nd annual conference of the cognitive science society, New Jersey: Erlbaum, (2000).

[11] S. Kaski, "Dimensionality reduction by random mapping: Fast similarity computation for clustering," In Proceedings of the IJCNN'98, International Joint Conference on Neural Networks. IEEE Service Center (1999).

[12] T. Landauer, and M. Littman, "Fully automatic cross-

language document retrieval using latent semantic indexing," In Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research, (1990) pages 31–38, Waterloo, Ontario, October.
[13] K. Lund, C. Burgess, "Producing high-dimensional semantic spaces from lexical co-occurrence," Behavior Research Methods, Instrumentation, and Computers, (1996) 28, 203- 208.
[14] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala, "Latent semantic indexing: A probabilistic analysis" In Proceedings of the 17th ACM Symposium on the Principles of Database Systems. ACM Press (1998).
[15] T. A. Plate, "Holographic reduced representations," IEEE Transactions on Neural Networks, 6, 623– 641, (1995).
[16] G. L. Recchia, M. N. Jones, M. Sahlgren, and P. Kanerva, "Encoding sequential information in vector space models of semantics: Comparing holographic reduced representation and random permutation," In S. Ohlsson and R. Catrambone (Eds.), Proceedings of the 32nd Cognitive Science Society, 865-870, (2010).
[17] M. Sahlgren, and R. Coster, "Using bag-of-concepts to improve the performance of support vector machines in text categorization," In Proceedings of the 20th International Conference on Computational Linguistics, COL ING'04 (pp. 487{493) (2004).
[18] M. Sahlgren, "An Introduction to Random Indexing," In Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE, Copenhagen, Denmark,(2005).
[19] M. Sahlgren, "The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces," Ph.D. Dissertation, Department of Linguistics, Stockholm University (2006).
[20] M. Sahlgren, "The Distributional Hypothesis. From context to meaning," Distributional models of the lexicon in linguistics and cognitive science (Special issue of the Italian Journal of Linguistics), Rivista di Linguistica, volume 20, numero 1, (2008).
[21] M. Sahlgren, A. Holst, and P. Kanerva, "Permutations as a Means to Encode Order in Word Space," Proceedings of the 30th Annual Meeting of the Cognitive Science Society (CogSci'08), July 23-26, Washington D.C., USA, (2008).
[22] G. Salton, "The Smart Retrieval System – Experiments in Automatic Document Processing," Prentice-Hall, Englewood Cliffs, NJ, 1971.
[23] G. Salton and M. McGill, "Introduction to modern information retrieval," McGraw-Hill, New York, NY, 1983.
[24] H. Schutze, "Automatic word sense discrimination," Computational Linguistics (1998) 24(1):97–124.
[25] A. Toumouh, A. Lehireche, D. Widdows, and M. Malki, "Adapting WordNet to the Medical Domain using Lexicosyntactic Patterns in the Ohsumed Corpus," In Proceeding of The 4th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA-06), Dubai/Sharjah, UAE,pp. 1029-1036, (2006).
[26] A. Toumouh, D. Widdows, and A. Lehireche, "Using Word Space Models for Enriching Multilingual Lexical Resources and Detecting the Relation Between Morphological and Semantic Composition," International Conference on Web and Information Tecnologies (ICWIT '08), pp. 195-201, (2008).
[27] A. Toumouh, D. Widdows, and A. Lehireche, "Parallel corpora and WordSpace models: using a third language as an Interlingua to enrich multilingual resources," International Journal of Information and Communication Technology, Vol. 3, No. 4, pp.299-313, (2011).
[28] T. Zesch, "Study of Semantic Relatedness of Words Using Collaboratively Constructed Semantic Resources," TU Darmstadt, Ph.D. Thesis, (2010).
[29] D. Widdows, "Geometry and Meaning," CSLI Publications (2004).
[30] D. Widdows, A. Toumouh, B. Dorow, and A. Lehireche, "Ongoing Developments in Automatically Adapting Lexical Resources to the Biomedical Domain," Fifth International Conference on Language Resources and Evaluation, LREC, Genoa, Italy, (2006).
[31] D. Widdows, and K. Ferraro, "Semantic vectors: a scalable open source package and online technology management application," The 6th Edition of the Language Resources and Evaluation, (LREC2008), Marrakech, Morocco (2008).
[32] R. Baeza-Yates, and B. Ribiero-Neto, "Modern Information Retrieval" Addison Wesley / ACM Press (1999).

Authors' Profiles

Adil TOUMOUH obtained his PHD in 2013 in the field of ontology learning. His research area includes the engineering of ontologies, computational linguistic, knowledge and web intelligence and NLP. He is a member of the Knowledge Engineering Team at the EEDIS laboratory. Dr. Adil TOUMOUH has many papers and has already contributed in the field of multilingualism with Prof. LEHIRECHE Ahmed and Dr. Dominic Widdows by proposing a method for the enrichment of multilingual resources using parallel corpora and algebraic model Word Space Model. The three authors contribute also on the field of adaptation of lexical-semantic resources to the biomedical domain. Currently he is a Teacher and researcher at the Institute of Computer Sciences of Liabess Djilali University, teaching object-oriented programming and information systems.

Dominic Widdows works principally on information extraction from the web for Bing local search. A mathematician by training, Dominic has worked on differential and algebraic geometry at Oxford (1996-2000), natural language processing and search at Stanford (2001-2004), distributed databases and collaboration at MAYA Design (2004-2007), and information extraction at Google and Microsoft Bing (2007-2015). His main theoretical research focus for a number of years has been on vector models for learning and reasoning, and the interaction between this area and quantum theory. He continues to contribute research papers in a range of areas including quantum informatics and concept learning, and works on several program committees and review panels. As well as being part of the Google Sky Map team, his main contributions to open source projects have been in the areas of semantic mapping and semantic search, including the Semantic Vectors package, initially created in partnership with the University of Pittsburgh, and now maintained by a small group of researchers

and developers internationally.

Ahmed LEHIRECHE has completed respectively ING Diploma from ESI of Algiers (1981) with the final curriculum project at the IMAG (France),"MAGISTER" Diploma from USTOran (1993) and "DOCTORAT D'ETAT" Diploma from UDL Sidi bel Abbes (2005). He is working as a Director of research, head of the Knowledge Engineering Team at the EEDIS laboratory and full Professor at the computer science department of UDL Sidi bel Abbes. He is mainly concerned with AI, Computer Science Theory and Semantics in IT.

How to cite this paper: Adil Toumouh, Dominic Widdows, Ahmed Lehireche,"Exploring Semantic Relatedness in Arabic Corpora using Paradigmatic and Syntagmatic Models", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.8, No.1, pp.37-47, 2016. DOI: 10.5815/ijieeb.2016.01.05

