Bassam Haddad*

Cognitively Motivated Query Abstraction Model Based on Associative Root-Pattern Networks

https://doi.org/10.1515/jisys-2017-0549
Received October 18, 2017; previously published online September 19, 2018.

Abstract: This paper attempts to investigate some aspects related to problems involved in textual inter-cognitive communication in the context of search queries. Furthermore, it aims at stressing on the root-pattern and morpho-phonetic dimension of a word meaning within a query, and its effects on understanding and predicting the intended information conveyed by some search patterns in a human language. As humans are inclined to use very few words possibly pervaded with vague and uncertain interpretational potential for requesting information, misinterpreting conveyed information in a query term might critically influence an inter-cognitive communication, particularly in case of Arabic- and Semitic-based computer systems. Furthermore, as phonetic patterns are involved in the mental word perception, an abstract morpho-phonetic query model is proposed based on the non-linearity of the morpho-phonetic characteristic of Arabic word cognition. This model suggests forming the intended query information by constructing morpho-phonetic query patterns relying on the most associative root-pattern subnetworks. An important advantage of this model resides in introducing the concept of query abstract morpho-phonetic vectors expressing query vector space. Furthermore, this approach suggests employing the fuzzy subsethood theorem as an assessment reflecting model accuracy and the closeness to human associative word-networks. Finally, it opens the discussion to consider indexing based on a higher level of abstraction, such as utilising patterns as cognitive search variables. Furthermore, as this model is capable of predicting most human associative query key terms, integrating these within certain human-machine interaction would improve inter-cognitive communications.

Keywords: Query expansion, intelligent information retrieval, root-pattern-based search, human-computer interaction, semantic representation, associative word-network, statistical language model, associative root-pattern networks, fuzzy subsethood theorem.

2010 Mathematics Subject Classification: 91EXX

1 Introduction

Humans tend to rely on the multiplicity provided by natural languages in expressing their request for information. A search topic or query can be formulated in different morphological forms and terms. Furthermore, humans are disposed to use a limited number of keywords for requesting information. For example, according to Klink et al. [20], about two to three words are the average term words used in a search query.

However, incomplete and ambiguous query words pervaded with uncertainty for requesting information form a potential source for misinterpreting the conveyed information in a query, particularly in case of Arabic and Arabic script-based computer systems. This problematic issue of natural languages has generally been investigated in cognitive infocommunication science [3], cognitive informatics and information retrieval (IR). Retrieving the most relevant documents within a large collection satisfying the intended human inquiry expressed in a small number of words is still a demanding task in IR and cognitive linguistics. Furthermore, using natural language understanding systems, as tools for inter-cognitive infocommunications

***Corresponding author: Bassam Haddad,** Department of Computer Science, University of Petra, Amman 11196, Jordan, e-mail: haddad@uop.edu.jo

between humans and machines, which are capable of simulating query cognition; represents an important step toward supporting the co-evolution process between human and machines.

In this framework, as Arabic is known for its hugely inflectional morphology, and its propensity to polysemy, on the stem, root, particle and pattern levels, the human-computer interaction can be harmfully affected. Incomplete and noisy query terms possibly spread out with real word spelling errors, complicate the inter-cognitive communication by recalling irrelevant and possibly hindering relevant information of becoming visible, depending on the employed indexing strategy, i.e. root-, stem- or lemma-based. Such types of miscommunication can be considered as a type of misunderstanding in an inter-cognitive communication between human and machines.

The following examples might illustrate the particularity of these issues in the case of Arabic:

Example (a): Although the following words share the same *abstract three-literal root*, i.e. (/*ktb*/) (the basic meaning of "writing") and even the same *stem*, (/*kitāb*/), the morphologically related word (/*kitāb*/, *Book*) is not relevant from the IR point of view:

- (/*kitābah*/, *Writing*).
- (/*kuttāb*/, *Writers*).

Example (b): A similar problem resides in the query (/*ʼlʻāmil ğaydun*/), which might be interpreted as (*the good factor*) or (*the good worker*), as both interpretations share the same stem, root and lemma.

Example (c): Even a cognitive phonetic spelling error [14] for one term query such as (/*quḍạt*) for (/*quḍah*/, *Judges*) might direct an IR process or a communication process towards another natural language using Arabic script for indexing, such as Persian. Furthermore, as the context of this model is cognitively motivated, employing such a biased communication in the context of inter-cognitive communication might be misleading.

Example (d): Furthermore, a query containing a semantic error, such as

(/*kuryāt ʼlḏm*/, *rebuking beads*),

would fail to retrieve even some relevant documents, as both query key-terms are real words, i.e. not non-words; however, replacing (ʼ*lḏm*/, *rebuking*) with (ʼ*ldm*/, *Blood*) within a query reformulation into

(/*kuryāt ʼldm*/, *blood beads*),

or an expansion into

(/*kuryāt ʼldm ʼlẖmāʼ*/, *red blood beads*)

would be a good attempt to predict the intended query information, which would improve the inter-cognitive communication between human and computer. However, the Arabic term for blood cells in the context of blood, such that (/*kurayāt*/, *beads*), can be misleading in the context of multi-lingual search, so that a query reformulation might be desirable in replacing it into (*ḫalāyā*, *Cells*):

(/*ḫalāyā ʼldm ʼlẖmrāʼ*/, *red blood cell*).

A major source for the occurrence of these difficulties might be explained by following observations.

Non-linearity of the morphology. Arabic and Semitic languages are dominantly non-linear (i.e. non-concatenative), whereas most indexing, search and query expansion techniques are based on a concatenative word structure such as Latin-based languages. Resolving this aspect within a search process might require considering the following:

- Different central elements of Arabic morphology such as roots, stems and lemmas. We propose even to consider patterns as part of the indexing process.

– Morpho-based contextual semantic word forms.

Root-pattern particularity in word perception. The current research concerned with cognitive word identification and recognition for Semitic languages, and in particular for Arabic and Hebrew, is increasingly providing evidence for the tendency of Arabic of being root-morpho-phonetic pattern based during visual and textual word processing and recognition [4, 7, 12]. Employing this principle in exploring a query potential provides us with a cognitively motivated approach for query expansion and resolving possible ambiguity working on a higher level of abstraction.

In this study, we propose a novel approach considering the above-mentioned observations relying on associative bi-directional relations introduced in the APRoPAT (Associative Probabilistic Root PATtern) model in the framework of query abstraction and construction [12]. This model can be regarded as one of the application potentials of the APRoPAT model in the context of human-centred interaction. Furthermore, as this model relies on semantic network representation, it is capable of supporting the core computational problems such as predication, disambiguation and gist extraction [25]. However, it differs from the classical associative word-network in the following sense:

– This model dominantly operates a higher level of abstraction than simple word-associative networks. In this context, roots are not real-world words; they are rather higher semantic abstractions of basic meanings of possible Arabic words and they are unalterable. Whereas, patterns are variables representing phonetic templates for some possible word forms and do not exist as real words. We can consider them as some kind of template of acoustic patterns representing the tones of possible words, which need roots to be recognised [4, 12, 16].
– The associative relationship between the abstract concept is always bi-directional to enable backward and forward chains.

The principal idea of the proposed model is based on proposing the most associative root-pattern subnetwork from query terms. Roots, stems, particles and patterns are defined as abstract cognitive variables on the morpho-phonetic level of word cognition. Based on extraction of certain associative relationships within these cognitive variables, a query morpho-phonetic network expressed in terms of degrees of association strength-function is established. Moreover, the notion of query morpho-phonetic vector is used to represent the most associative query vector within the query vector space. The initial motivation relies on the characteristic of human language cognition, particularly in Semitic languages; that is, phonetic patterns are involved in the mental word perception and identification, which can be expressed in terms of the morpho-phonetic characteristic of Arabic in the context of query perception.

1.1 Non-Linearity of Morphology and Word Cognition

Cognitively, a language can be separated into various linguistic levels, e.g. phonology, morphology, syntax and semantics. However, the minimal linguistic levels required for word cognition is controversial. Cognitive grammar depends on two levels, i.e. *phonology* and *semantics* with symbolic links between them, in the sense that all linguistic units are symbolic units relating a meaning directly to a phonological form [8]. Word perception in Arabic appears to treat these constituents in a distinctive form. Phonology and morphology are mainly non-linear. This aspect complicates the problem about the number and order of the mediated cognitive levels involved in perception of a word, as different levels might be non-linearly involved in a word perception. The current research concerned with cognitive aspects of Semitic languages testifies constantly that “Semitic languages and in particular Arabic are strong root-patterns” [4, 7]. However, this aspect is still debatable in the Arabic computational community. According to Ref. [12], a word form is based on of the following characteristics:

– Root-tier mostly consisting of three consonants representing the highest autonomous semantic abstraction and it is changeless. An associative network of roots represents an abstract semantic class without obvious phonetic information.

- Templatic pattern in the form of consonant-vowel concatenation describing the morpho-phonetic form of a word integrating some possible phonetic, syntactic and semantic information.
- Root-pattern association. A pattern variable can be sensed by building the most conceivable, associative relationship between a pattern and a root [11]. In terms of lambda-abstraction, this process can been expressed as an applicative function of a root to a phonetic pattern:

$$\lambda\langle \mathbf{R_i}\rangle_{\in \mathrm{ROOT}} \cdot \left[\langle \mathbf{Pt_j}\rangle\right]. \tag{1}$$

1.2 Related Work

The Arabic natural language processing research community has unfortunately set little efforts in discussing this issue from a cognitive point of view. However, in the context of IR research, there are some reports addressing this issue in connection to query expansion and relevance feedback. Many of these approaches are increasingly considering *term dependency* metrics such as "word proximity" and "word co-occurrences" [26]. However, few of these techniques are motivated from a cognitive viewpoint, besides the fact that the term of *relevance* is still a relative measure in IR [10, 13]. In this context, there is some research treating this aspect relying on interactive "word sense disambiguation" by expanding some term queries towards resolving some possible term polysemy by selecting interactively not ranked, but more specific synonyms [2]. There are also some reports treating query expansion in terms of word proximity and coherence [18] in the form of considering *terms similarity* using expectation maximization [24], and building automatic Arabic thesauri using *term-term similarity* [19]. In addition, there is also some research utilising *ontologies* such as Arabic WordNet [1, 23], and root extraction considering thesauri for query expansion [22]. This approach is based on extracting the best WordNet synset concepts to act as source for further query expansion, and it is similar to our model; however, our approach is focused on extracting the most associative query subnetwork, besides the fact that our model is capable of deriving the implicit synsets from the global APRoPAT associative network.

In addition to studies about cognitive word identification and mental lexical representation, there is, in the meanwhile, some research concerned with cognitive word identification in the context of visual and textual word processing. These models rely on the assumption that word cognition can be considered as a dynamic process operating on the mental lexicon trying to identify a word by establishing an associative pattern-root and root-pattern relationship to perceive the phonetic structure of the word. This aspect has been adopted by considering the bi-directional characteristic of the APRoPAT dataset [17]. The employed dataset itself contains around 11.5 billion word forms and 9.3 million associative relationships. Furthermore, our approach is also similar to some extent to that of Booth et al. [6] and Deb Roy and Zeng [9] in the sense that our approach considers the construction of a query subnetwork as a basis for query recovery and reformulation; however, it relies on the cognitive concepts of the APRoPAT dataset employing more specific relations such as associative bi-directional root-pattern relations, and on composing new associative relations based on the compositional rule of inference.

1.3 Motivation and Scope of the Paper

This paper focuses the attention on some characteristics of a cognitively motivated query construction model relying on global "associative bi-directional binary relationships" [12]. Details of the potential application and complexity of the APRoPAT model lie beyond the scope of this presentation.[1]

This study is furthermore not intended to introduce a complete model for query lexical semantics or a query expansion model in the classical sense to be applied in context of a certain IR model; it aims

1 The application potential of the APRoPAT model is wide ranging, and it has already been employed in developing non-word detection and correction system [14] and improving the Petra-Morph Morphological Analyser. At present, we are working on a multi-modal search engine considering indexing on different levels, such as root, stem, lemma and root-pattern, besides integrating this model within an query expansion interface: http://apropat.info/portal/apropat-search-engine/.

rather at stressing the importance of using a global bi-directional associative semantic network in query construction and the meaning of the morpho-phonetic dimension in predicting query forms, which might be close to human query formulation and understanding. However, integrating such models within certain human-machine interactions such as IR and inter-cognitive communication systems would be an important improvement in query expansion modelling.

Furthermore, the fuzzy-subsethood theorem [15, 21] will be proposed to evaluate this model in terms of similarity and closeness to a human-based word associative test collection; therefore, this presentation it is not intended to measure the performance of some specific indexing methodology within a query expansion model in the context of a specific IR system.

The remainder of this paper is structured as follows. The formal definition of the notion of ⟨*cognitive variable*⟩ – *headed subnetwork*, *local query associative subnetwork* and the *abstract phonetic vector of a query* will be introduced, followed by a model evaluation experiment.

2 Proposed Processing Model

As stated earlier, the proposed model is cognitively motivated and human centred, which proceeds from the view that a query, textual, contextual and conceptual knowledge can be extracted from an associative bi-directional semantic network established between basic cognitive variables on the morpho-phonetic and semantic levels of perception. Furthermore, a query can consequently be represented as a vector of keywords going together with a certain phonetic search information. Operating on roots and patterns leverage associative relationships on a higher level of abstraction. An associative root-root relationship implies multiple word-word co-occurrences (see Figure 1). Subsequently, a query represented by basic cognitive variable such as ROOT, STEM, PARTICLE and PATTERN can be expanded into a larger associative bi-directional network representing a local knowledge of the involved query. Roots and pattern are decisive in extracting textual and contextual query terms. Extracted words can be classified as *textual* (i.e. based on the root literal) and *contextual* (i.e. based on associative root-pattern relationships).

The process of anticipating the intended users' requests can be understood as a semantic process directed toward building the most associative similar query root-pattern subnetwork. This process implicates instantiating the most reasonable query vector within the morpho-phonetic space of a query. Figure 2 shows the architecture of the involved processes:

- Morphological data processing;
- Query subnetwork construction;
- Morpho-phonetic query vector construction;
- Query instantiation;
- Query candidate generation.

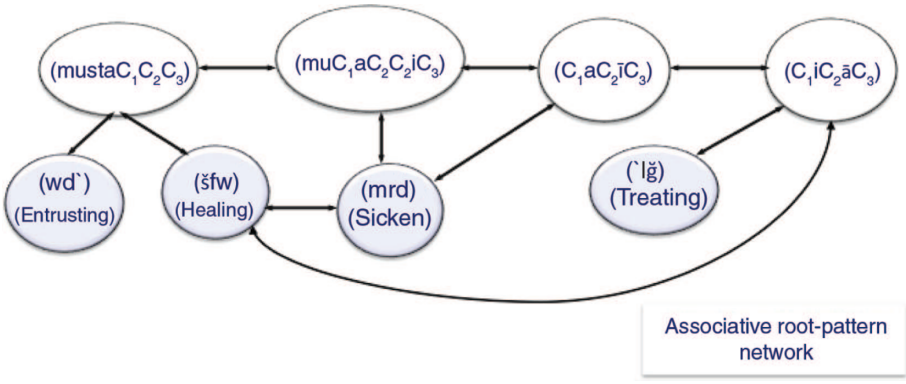

Figure 1: An Example of Associative Root-Pattern (Morpho-Phonetic) Network Containing Bi-directional Associative Relationship between Roots and Pattern.
The pattern (/*mustaC*$_1$*C*$_2$*C*$_3$/) might sound phonetically like (/*mustašfā*/, *Hospital*); however, it is cognitively not perceivable as long it has been not applied to the root (/šfw/ *healing*).

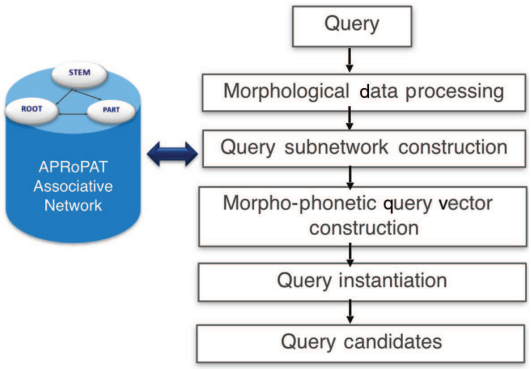

Figure 2: Process of Construction of an Abstract Morpho-Phonetic Query Vector Based on Associative Cognitive Variables Relationships and Their Instantiation.

Table 1: Considered Root Pattern.

Dataset items	Size
ROOT	8065
PATTERN	5737
PARTICLE	281
Non-linear word forms	11,322,272

Some major benefits of this model can be summarised as follows:

- Pattern polysemy and root homonymy can also be decreased by considering the most associative relative query root-pattern vector.
- New contextual and textual associative word forms can also be extracted.

Next, we will focus attention[2] on the most relevant relations in the context of local query root-pattern network construction.

Definition 1 (model cognitive variables): A morpho-phonetic cognitive variable $\mathcal{X}$ on the level of morpho-phonetic level of cognition is a type of the following categories:

- ROOT, denoting the set of all in the model-identified roots.
- PATTERN, denoting the set of all in the model-identified patterns.
- STEM, denoting the set of all in the model-identified stems.
- PARTICLE, denoting the set of all in the model-identified particles.

The set of all identified variables is subject to APRoPAT dataset construction [17]. However, Table 1[3] gives an overview of some identified variables.

Definition 2 (associative relationships): Let $\mathfrak{M}$ be the set of all (in the model involved morpho-phonetic) cognitive variables, and $\mathcal{X}$, $\mathcal{Y} \in \mathfrak{M}$, then a binary associative relationship $\mathcal{X} - \mathcal{Y}$ between $\mathcal{X}$ and $\mathcal{Y}$ is defined as

$$\mathcal{X} - \mathcal{Y} \subset \mathcal{X} \times \mathcal{Y}, \tag{2}$$

associated with some **CO**gnitive **D**egree of **A**ssociation **S**trength-Function

$$\textbf{CODAS}(\mathcal{X}, \mathcal{Y}) \text{ for each variable} \in \mathcal{X} - \mathcal{Y}, \tag{3}$$

where the cognitive degrees of association is annotated by

$$P(\mathcal{X}|\mathcal{Y}), \forall\, x_i \text{ and } y_j \in \mathcal{X} \text{ and } \mathcal{Y}, \tag{4}$$

2 Details of morphological data processing and query candidate generation are out of the scope of this presentation.

3 Based on Petra-Morph, designed and implemented by Arabic TEXTWARE Company and University of Petra Team [17].

respectively, by considering a large-scale corpus. Analogically, the other direction $\mathcal{Y} - \mathcal{X} \subset \mathcal{Y} \times \mathcal{X}$ can be constructed, $\forall \mathcal{X}$ and $\mathcal{Y} \in \mathfrak{M}$.

Based on Definition 2, a global associative network on the morpho-phonetic level can be viewed as a network of all bi-directional associative relationships between all involved morpho-phonetic cognitive variables $\in \mathfrak{M}$. Furthermore, the strength of association between these binary relations can also be expressed in terms of binary fuzzy relations, whereas the degree of the strength of association is estimated through a corpus considering bi-directional prior probabilities.

To generalise and abstract the meaning of association, we will merely use the $CODAS(\mathcal{X}, \mathcal{Y})$ to annotate some function determining the grade of association between cognitive variables. Moreover, $\forall\ x_i$ and $y_j \in \mathcal{X}$ and $\mathcal{Y}$, respectively, we use $codas(x_i, y_j)$ to formalise the instantiated (concrete) degree of strength of association between some instances x_i and y_j in the global network. In the proposed model, APRoPAT-based associative relations were adopted as an estimation for $CODAS(\mathcal{X}, \mathcal{Y})$.

Example 1: Let (*'isti***C₁C₂ā C₃**) be an instance ∈ PATTERN and the three radical root (/**'ml**/, *Working*) ∈ ROOT, then functionally applying the abstract root [11] (/**'ml**/, *Working*) to the pattern template instance (*'isti***C₁C₂ā C₃**) generates a new word form (/*'isti'māl*/, *Utilization*). However, the degree of the associative relationship between the root (/**'ml**/, *Working*) and the cognitive variable of being the concrete pattern (*'isti***C₁C₂ā C₃**) is annotated by the estimated associative value, **codas**('ml,' *isti***C₁C₂ā C₃**), which can be estimated by calculating the priori such as **P**('*isti***C₁C₂ā C₃** | **'ml**):

$$\langle (\textbf{/'ml/}, \textit{Working}) \rangle \xrightarrow[\textbf{codas}(\text{'}ml,\text{'}isti C_1 C_2 \bar{a} C_3)]{} \langle \text{'}isti\mathbf{C_1 C_2 \bar{a} C_3} \rangle \tag{5}$$

On the other hand, understanding the basic meaning of a phonetic pattern such as the word form ('*isti***C₁C₂ā C₃**) is plausible under considering some lexical root such as (/**'ml**/, *Working*) by relying on its degree of association strength:

$$\langle (\textbf{/'ml/}, \textit{Working}) \rangle \xleftarrow[\textbf{codas}(\text{'}isti C_1 C_2 \bar{a} C_3, \text{'}ml)]{} \langle \text{'}isti\mathbf{C_1 C_2 \bar{a} C_3} \rangle \tag{6}$$

Regarding a root as the highest level of symbolic semantic abstraction permits the recognition of words as bi-directional probabilistic applicative process for instantiating the most reasonable pattern to the most perceivable root, in an associative network between cognitive variables.

In terms of lambda-abstraction, this process can been expressed as an applicative function of a root to a phonetic pattern:

$$\lambda \langle \mathbf{R_i} \rangle_{\in \text{ROOT}} . \left[\langle \mathbf{Pt_j} \rangle \right] \tag{7}$$

with $codas(R_i, Pt_j)$ as possible strength degree of association.

This expression means to find or search for an associative root in the mental lexicon, i.e. the (most plausible) associative or consistent root in the context of a given morpho-phonetic pattern $\mathbf{Pt_j} \in$ PATTERN. More specific than the above example:

$$\lambda \langle (\text{'}\mathbf{istiC_1 C_2 \bar{a} C_3}) \rangle_{\in \text{PATTERN}} . \left[\langle (\textbf{/'ml/}, \textbf{Working}) \rangle \right]$$

with *codas*(/'*ml*'/, *'isti* C_1 $C_2\bar{a}$ C_3) as possible strength degree of association.

Analogically, the process of generating a word can be understood as a process of finding an associative morpho-phonetic pattern under thinking of some abstract semantics represented by an abstract root such as $\mathbf{R_i}$, e.g. (/'ml/, Working):

$$\lambda \langle \mathbf{Pt_i} \rangle_{\in \text{PATTERN}} . [\langle \mathbf{R_i} \rangle] \tag{8}$$

with $codas(Pt_j, R_i)$ as possible strength degree of association.

This expression means to find (the most plausible) associative or a consistent pattern form for a given root. By further analysis, this process can be understood as an uncertain rule between the given root and a possible pattern.

2.1 Cognitive Variable-Headed Associative Network

As indicated earlier, each instance of a cognitive variable is interconnected by multiple bi-directional associative relationships. However, considering a query as a network of related morpho-phonetic forms allows us to follow the potential semantics of a query by determining the most associative textual and contextual subnetwork for each word form of the query. Based on a global associative network, a local associative subnetwork for each reduced query word can be constructed. Subsequently, query morpho-phonetic vectors can be followed by determining different query *strong associative morpho-phonetic paths*, where each vector can be considered as a query expansion model for the involved query words.

At this level, a query subnetwork represents an abstract morpho-phonetic model for cognition, where roots and their possible pattern variables are processed bi-directionally to determine concrete word candidates required to perceive a word.

Next, we will focus on definitions of cognitive variable-headed instances as subnetworks, followed by the meaning of the concept of abstract morpho-phonetic vectors and some characteristics of the model.

Definition 3 (cognitive variable-headed associative network): Let $\mathbf{x}$ be instance for some cognitive variable and $\mathbf{Y}$ be any cognitive variable $\in \mathfrak{M}$, then the $\langle \mathbf{x} \rangle$-headed associative subnetwork is defined as the sub-graph associated with the $\mathbf{x}$-Y relation for "$\mathbf{x}$" and $\forall\, y_i \in \mathrm{Y}$ and $\forall$ **codas**(x, y_i), where

$$\mathbf{x} - \mathrm{Y} \triangleq \{((x, y_i), \mathbf{codas}(x, y_i)) \mid (x, y_i) \in \mathrm{X} \times \mathrm{Y}\} \tag{9}$$

The significance of Definition 3 comes to effect when considering a query involving terms in their reduced form, such as roots, patterns, stems or particles. Focusing the search process on certain central query term subnetworks directs the constructing process towards proposing a query local associative subnetwork, where certain plausible query vectors can be predicted as candidates for query expansion or more precisely for completing missing contextual or textual terms in the original query.

Definition 4 (query associative root-pattern-local network): Let Q be a query, $Q = \{W_1, .., W_n\}$, containing the words $W_1, \ldots, W_n$ with the corresponding reduced cognitive instances, $w_1, w_2, .., w_n$. The related query network is constructed by establishing the most associative $\langle \mathbf{w_i} \rangle$-headed subnetworks of the involved words.

In Figure 3, the instances $\mathbf{R_1}$ and $\mathbf{Pt_3}$ represent an *abstract textual* and *contextual* associative expansion of the original query root, i.e. $\mathbf{R_w}$. The cognitive degree of association strength is depicted by the *CODAS*(X,Y) function on the morpho-phonetic level of abstraction, e.g. $codas(R_w, Pt_3) \in [0.75, 1]$. Applying $\mathbf{Pt_3}$ to root $\mathbf{R_w}$ in the form of $\langle\langle \mathbf{Pt_3} \langle \mathbf{R_w} \rangle\rangle$ would generate a new word form with the same basic semantic of the query root $\mathbf{R_w}$. Figure 4 represents three local associative subnetworks of a query consisting of three words, i.e. $\mathbf{Q} = \langle \mathbf{W_1}, \mathbf{W_2}, \mathbf{W_3} \rangle$, whereas Figure 5 shows the resulting associative subnetwork of the same query.

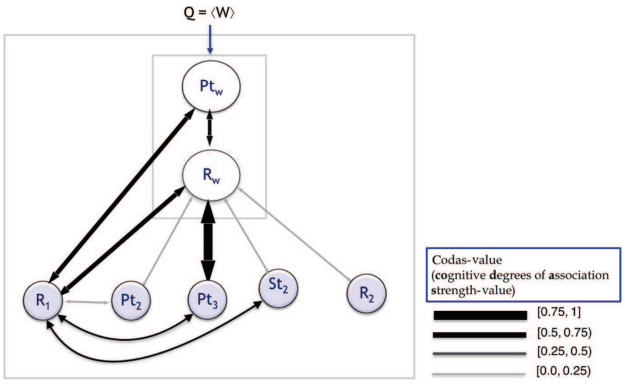

Figure 3: A Local Associative Subnetwork for a Query Consisting of One Word Query Term, i.e. $\mathbf{Q} = \langle \mathbf{W} \rangle$.
The query is headed by the instance pattern $\langle Pt_W \rangle$ and the root instance $\langle R_W \rangle$. The instances $\langle R_1 \rangle$ and $\langle Pt_3 \rangle$ represent a possible contextual one-term query candidate for the word **W**. The degree of association strength is depicted by intervals for simplification.

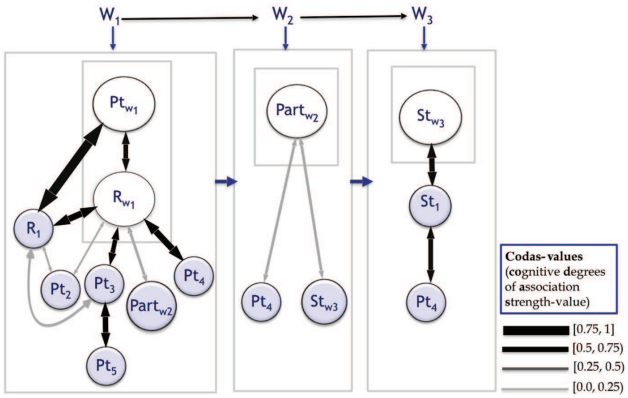

Figure 4: Local Associative Subnetworks of a Query Consisting of Three Keywords: $Q = \langle W_1, W_2, W_3 \rangle$.

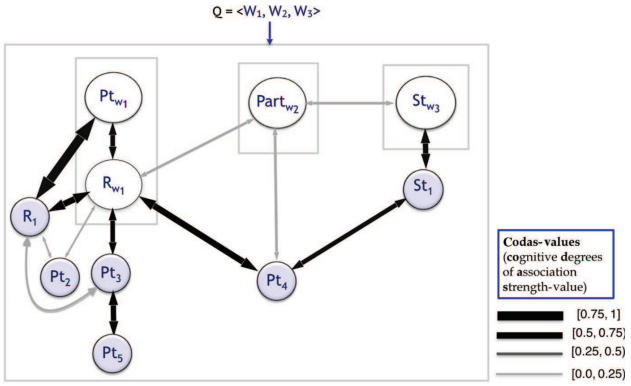

Figure 5: A Joint Local Associative Subnetwork for a Query Consisting of Three Words: $Q = \langle W_1, W_2, W_3 \rangle$.

2.2 Query Morpho-Phonetic Vectors

Based on the APRoPAT model, an instance of a cognitive variable has multiple associative links with other instances, in the form of associated bi-directional relationships represented by a CODAS-Function, e.g. a root with multiple patterns. Certain instances, especially patterns, appearing explicitly in the query have an important effect in building the phonetic vector of the query. Cognitively, a query phonetic vector represents the basic phonetic knowledge of the intended information in the form of templatic syntactical and semantic abstract data. The phonetic vector of a query is cognitively not fully conceivable without instantiating it with the most plausible root instances. Furthermore, considering a query phonetic space within a local associative network provides a preliminary source for expanding and completing query textual and contextual relationships.

In the following, the definition of query abstract morpho-phonetic vector will be introduced.

Definition 5 (abstract morpho-phonetic vector of a query): Let Q be some query, $Q = \langle W_1, ..., W_n \rangle$, containing the words W_1, ..., W_n with the corresponding vector of the reduced instances, $\langle w_1, ..., w_n \rangle$. A query morpho-phonetic vector is defined as the most associative abstract template of cognitive variables extracted based on the involved binary associative relations within the query associative subnetwork.

2.2.1 Composing Morpho-Phonetic Query Vector Variables

An abstract phonetic vector consists of a linear structure of strong associative instances of certain cognitive variables extracted from the reduced query terms and words. At this instance, it is worthwhile to mention

that depending on the type and frequency of the involved key terms, the corresponding local query subnetwork can be of high dimensionality. Establishing a query network with strong associative relations expressed by their associative strength would reduce the dimensions of a targeted query vector to serve as a template for instantiation of different query forms. However, this paper proposes capturing the concept of strong associative cognitive variables by:

- The maximal local paths in a query subnetwork;
- Composing and inferring new associative relations by employing a variation of MAX-MIN compositional rule of inference, for composing new suggestions for strong associative terms on the morpho-phonetic level of cognition, e.g. point valued and interval valued.

Definition 6 (composing strong associative cognitive variables): Let $\mathfrak{M}$ be the set of all (in the model involved) morpho-phonetic types and $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z} \in \mathfrak{M}$, and $\mathcal{R}_1 = \mathcal{X} - \mathcal{Y}$, $\mathcal{R}_2 = \mathcal{Y} - \mathcal{Z}$ are binary associative relations, then a strong associative relationships is defined as

$$\mathbf{codas}(\mathbf{x_k}, \mathbf{z_j})_{\mathcal{R}_1 o \mathcal{R}_2} = \underset{y_i \in \mathcal{Y}}{max\, min}\{codas(x_k, y_i)_{\mathcal{R}_1}; codas(y_i, z_j)_{\mathcal{R}_2}\}. \tag{10}$$

Furthermore, by direct relations, the most associative relation is estimated as the maximal argument $codas(x_i, y_j) \in \mathbf{CODAS}(\mathcal{X}, \mathcal{Y})$.

For example, based on Figure 5, the following morpho-phonetic query vectors can be composed:

(a) $\langle\langle Pt_{w1}\, \langle R_{w1}\rangle\rangle, Part_{w2}, St_{w3}\rangle$;
(b) $\langle\langle \mathbf{Pt_4}\, \langle R_{w1}\rangle\rangle, \mathbf{St_1}\rangle$;
(c) $\langle\langle Pt_{w1}\, \langle \mathbf{R_1}\rangle\rangle, \langle \mathbf{Pt_4}\, \langle R_{w1}\rangle\rangle, \mathbf{St_1}, St_3\rangle$.

The query template in (a) consists of an abstract pattern and its root, i.e. $\langle Pt_{w1}\langle R_{w1}\rangle\rangle$, followed by the query original particle and stem. However, in (b), the original root in w_1, i.e. R_{w1}, is proposed to be expressed in a new phonetic pattern $\langle \mathbf{Pt_4}\langle R_{w1}\rangle\rangle$, which means the same basic semantic but within a new word form.

The following example (Example 2) illustrates the process of creating concrete queries based on the concept of the abstract phonetic vector of a query. It consists of two phonetic patterns followed by a particle and a stem.

Example 2: The following query structure consists of four word forms. It contains two patterns with their roots, followed by a particle and a stem. It represents an example of concretised abstract phonetic vector of a query instantiation:

$$\mathbf{Q} = \langle\langle \mathbf{C_1 i\, C_2 \bar{a} C_3 ah}\langle R_1\rangle\rangle, \langle \mathbf{C_1 i\, C_2 \bar{a} C_3}\langle R_2\rangle\rangle, \langle \mathbf{Particle}\rangle, \langle \mathbf{Stem}\rangle\rangle,$$

applying the following reduced query instance:

$$(\mathbf{/drs/}, learning), (\mathbf{/gm`/}, gathering), (\mathbf{/f\bar{i}/}, in), (/^{'}\mathbf{\bar{u}r\bar{u}ba/}, Europe),$$

it would produce the following query instance:

$$\begin{aligned} Q^{'} &= \langle `ldirasah, `lg\bar{a}mi`yah, f\bar{i}, {}^{'}\bar{u}r\bar{u}ba\rangle \\ Q^{'} &= \langle the\ academic, studying, in, Europe\rangle. \end{aligned}$$

3 Model Evaluation

Evaluating a generic model based on subnetwork abstraction and construction is difficult to standardise and formalise. In this model, the main reason for this difficulty relies on the following facts:

- The major outcomes of this model are query subnetworks, i.e. weighted graphs and templates for abstract phonetic vectors, which are difficult to evaluate with traditional evaluation methods.

- The associative values between concepts are based on fuzzy binary relations, which need fuzzy assessments.
- A major goal of this model is to stress on the cognitive aspect of this model and its closeness to human behaviour in the context of topic expansion. Measuring the performance in terms of IR metrics is out scope of this presentation at this stage of research.

To consider these aspects, this approach proposes employing the fuzzy subsethood theorem [21] as a metric to highlight the major aspects of the model:

- Representing constructed query associative word networks (M) as fuzzy subsets expressing the predicted concepts in relationship to the involved test queries, i.e. the outcome.
- Human-based associative word networks (H) represented as a fuzzy subset acting as standard and a reference for the assessment, i.e. the test gold standard.
- Bi-directional fuzzy subsethood degrees between M and H [15], i.e. $F\text{-}S(H,M)$ and $F\text{-}S(M,H)$ as evaluation metric assessing the similarity between human model and machine model:

$$F\text{-}S(H, M) \triangleq Grade(M \subset H), \text{ where} \tag{11}$$

$$F\text{-}S(H, M) \triangleq \frac{\sum_{i=1}^{n} Min\{\mu_M(x_i), \mu_H(x_i)\}}{\sum_{i=1}^{n} \mu_H(x_i)} \tag{12}$$

and analog

$$F\text{-}S(M, H) \triangleq Grade(H \subset M). \tag{13}$$

In this context, $F\text{-}S(H, M)$ is assessing the grade of the overall fuzzy-based accuracy of the model, where $F\text{-}S(M,H)$ is assessing the grade of the subsethood of the human-based associative model in the proposed query model. Here, it is worthwhile to mention that $F\text{-}S(H, M)$ is similar to the concept proposed by Binaghi et al. [5] as a fuzzy set-based accuracy assessment of soft classification.

3.1 Experiment

To the author's knowledge, there exists no standard dataset for the evaluation of Arabic query construction approaches based on the morpho-phonetic level of cognition. A cognitively motivated query relevance corpus has partially been utilised for evaluating the proposed query model. The original dataset was designed for creating relevance corpus containing around 110 topics and 1100 documents extracted from ClueWeb09, which contains around 29 million Arabic webpages. The relevance dataset contains 20,710 relevance assessments. The source of the topics was created based on a list of the top 1000 most frequent terms in the ClueWeb09 corpus. Words were manually chosen to be meaningful general terms. Details of this corpus are out of scope of this presentation; however, an overview of the generated data can be found in Ref. [13].

For creating human-based associative word-networks, 21 persons (predominantly university students) from different Arabic countries were requested to formalise, under the priming principals after appearing different types of topics for a short time, any priming terms without any reference to any text. Furthermore, later on and on another level, a query-related text document was available for the assessors and they were requested to reconstruct the original topic, i.e. to formulate a new query with their own words after disappearing the related document. Later on and on an advanced phase of the experiment, the assessors were requested to assess the grades of topic association in the generated word-networks by choosing values $\in [0, 1]$. Human-Based and Human-Document-based Associative word-networks (H-B-A and H-D-A, respectively) were created to act as standards for evaluating the accuracy of the proposed model in terms of $F\text{-}S(H, M)$ and human influence in the model by $F\text{-}S(M, H)$. To perform this experiment, a special interactive program

within a voting system was developed considering the overall term association agreement. The overall agreement among all assessors was estimated based on the average values among a group of judgments for each involved associative word-network.

Furthermore, for testing the proposed model, different types of queries were selected (one term, unstructured, structured queries) from the original H-B-A and H-D-A networks. A model-based query was regarded as correct to certain degree if its terms were visible in the corresponding H-B-A or H-D-A word-network of a test query.

Example 3: For illustration of this concept, let $q = (ḫalāyā, Cells)$ be a one-term query with the following H-fuzzy subset generated based on the human priming principal and evaluation given in the experiment. The values are given as point-valued fuzzy subset expressing the degrees of memberships, which can be interpreted as an associative word-network for the query:

$$H_{(ḫalāyā,Cells)} = \{(ǧaḏ‘, stem)/0.6, (blāzmā, plasma)/0.5, (naḥl, bees)/0.5, \\ (’rhāb, terror)/0.6(’nsān, human)/0.5, (šams, sun)/0.4, \\ (‘ṡab, nerve)/0.6\}$$

Furthermore, let $M_{(ḫalāyā,Cells)}$ be a fuzzy subset generated from the corresponding model query subnetworks and their abstract phonetic vectors:

$$M_{(ḫalāyā,Cells)} = \{(ǧaḏ‘, stem)/0.4, (blāzmā, plasma)/0.5, (naḥl, bees)/0.5, \\ (’rhāb, terror)/0.1,(’nsān, human)/0.6, (šams, sun)/0.5, \\ (‘ṡab, nerve)/0.7\}$$

Based on Eq. (12), the subsethood values of sets can be estimated as $F\text{-}S(H, M) = 0.81$ and $F\text{-}S(M, H) = 0.90$. It is obvious that the accuracy of the model is underestimating the value of (*’rhāb*, *terror*) with 0.1, and its association to the word (*ḫalāyā*, *Cells*). Otherwise, the rest of the values are close to the human assessment.

The fuzzy-based accuracy, i.e. $F\text{-}S(H, M)$, of the experiment considering 20 test queries of different types ranged from 0.79 to 0.92 with an average value of 0.83 and the closeness of human to model estimation, i.e. $F\text{-}S(M, H)$, ranged from 0.75 to 0.93 with an average of 0.87. These results support the view that the model is cognitively human centred.

4 Outlook and Conclusion

This paper attempted to introduce an elementary study towards establishing a novel approach concerned with some aspects related to formalising a cognitively motivated model for query abstraction. The proposed approach relies on establishing the most associative abstract phonetic templatic structures from local associative networks on the morpho-phonetic level of abstraction. The strength of association is expressed in terms of associative relations, which can be interpreted as binary fuzzy relations. To ensure the general and global character of this model, the degrees of the strength of the association were estimated based on a global corpus containing around 11.5 billion word forms and 9.3 million associative relationships. Some types of word polysemy and homonymy can also be reduced during constructing the corresponding most associative query subnetwork. A major goal of this presentation is to emphasise the morpho-phonetic and root-pattern dimension and its effects in understanding and predicting the intended information conveyed by some query patterns in a human language. This concept provides us with a novel approach to investigate a query and topics on morpho-phonetic level of abstraction. This aspect is important as this model is dealing with query understanding rather the internal indexing of an IR system. However, integrating such models within certain human-machine interaction such as IR and inter-cognitive communication systems would be an important

improvement in query expansion modelling. Expanding and integrating an IR index of into APRoPAT semantic is a subject of our future work. For the evaluation of this approach, the fuzzy subsethood theorem has been proposed as an assessment metric reflecting the fuzzy accuracy and the closeness to human-created associative word-networks. These associative word-networks were created under considering the priming effect before reading query terms and after reading certain related documents and human-based evaluation.

Bibliography

[1] L. Abouenour, K. Bouzouba and P. Rosso, An evaluated semantic query expansion and structure-based approach for enhancing Arabic question/answering, *Int. J. Inform. Commun. Technol.* **3** (2010), 37–51.

[2] R. Al-Shalabi, G. Kanaan, M. Yaseen, B. Al-Sarayreh and N. Al-Naji, Arabic query expansion using interactive word sense disambiguation, in: *The 2nd International Conference on Arabic Language Resources & Tools, MEDAR*, Cairo, Egypt, 2009.

[3] P. Baranyi, A. Csapo and G. Sallai, *Cognitive Infocommunications (CogInfoCom)*, Springer International Publishing, Basel, Switzerland, 2015.

[4] S. Bentin and R. Forst, Morphological factors in visual word identification in Hebrew, in: L. B. Feldman (ed.), *Morphological Aspects of Language Processing*, pp. 271–292, Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, USA, 1995.

[5] E. Binaghi, E. A. Brivio, P. Ghezzi and A. Rampini, A fuzzy set-based accuracy assessment of soft classifications, *Pattern Recognit. Lett.* **20** (1999), 948–935.

[6] J. Booth, B. Di Eugenio, I. F. Cruz and O. Wolfson, Query sentences as semantic (sub) networks, in: *Proceedings of the IEEE International Conference on Semantic Computing*, Berkeley, CA, USA, 2009.

[7] S. Boudelaa, Is the Arabic mental lexicon morpheme-based or stem-based? Implications for spoken and written word recognition, in: E. Saiegh-Haddad and R. M. Joshi (eds.), *Handbook of Arabic Literacy, Literacy Studies 9*, Springer Science+Business Media, Dordrecht, 2014.

[8] W. Croft and A. Cruse, *Cognitive Linguistics*, Cambridge University Press, Cambridge, UK, 2004.

[9] S. Deb Roy and W. Zeng, Cognitive canonicalization of natural language queries using semantic strata, *ACM Trans. Speech Lang. Process.* **10** (2013), Article 20, 30 pages.

[10] N. El-Khalili, B. Haddad and H. El-Ghalayini, Language engineering for creating relevance corpus, *Int. J. Softw. Eng. Appl.* **9** (2015), 107–116.

[11] B. Haddad, Probabilistic bi-directional root-pattern relationships as cognitive model for semantic processing of Arabic, in: *3rd IEEE International Conference on Cognitive Infocommunications 2012*, pp. 284–279, Kosice, Slovakia, 2012.

[12] B. Haddad, Cognitive aspects of a statistical language model for Arabic based on associative probabilistic Root-PATtern relations: A-APRoPAT, *Infocommun. J.* **V** (2013), 2–9.

[13] B. Haddad, Relevance & assessment: cognitively motivated approach toward assessor-centric query-topic relevance model, *Acta Polytech. Hung., J. Appl. Sci.*, Special issue on Cognitive Infocommunications, in press, 2018.

[14] B. Haddad and M. Yaseen, Detection and correction of non-words in Arabic: a hybrid approach, *Int. J. Comput. Process. Orient. Lang.* **20** (2007), 237–257.

[15] B. Haddad and A. Awwad, Representing uncertainty in medical knowledge: an interval-based approach for binary fuzzy relation, in: *The International Arab Journal of Information Technology, IAJIT*, Zarqa University, Jordan, 2010.

[16] B. Haddad, N. El-Khalili and M. Hattab, A cognitive query model for Arabic based on probabilistic associative morpho-phonetic sub-networks, in: *5th IEEE Conference on Cognitive Infocommunications*, Vietri sul Mare, Italy, 2014.

[17] B. Haddad, A. Awwad, M. Hattab and A. Hattab, Associative root-pattern data and distribution in Arabic morphology,*Data* **3** (2018), 10.

[18] E. Hoenkamp, Why information retrieval needs cognitive science: a call to arms, in: *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, Stresa, Italy, 2005.

[19] H. Khafajeh, G. Kanaan, M. Yaseen and B. Al-Sarayreh, Automatic query expansion for Arabic text retrieval based on association and similarity thesaurus, in: *Proceedings the European, Mediterranean & Middle Eastern Conference on Information Systems (EMCIS)*, Abu Dhabi, UAE, 2010.

[20] S. Klink, A. Hust, M. Junker and A. Dengel, Improving document retrieval by automatic query expansion using collaborative learning of term-based concepts, in: *Proceedings of the 5th International Workshop on Document Analysis Systems (DAS 2002) of Lecture Notes in Computer Science*, volume 2423, pp. 376–387, Springer, Princeton, NJ, USA, 2002.

[21] B. Kosko, *Neural Networks and Fuzzy Systems: A Dynamical Systems Approach to Machine Intelligence,* Prentice-Hall Inc., Upper Saddle River, NJ, USA, 1991.

[22] T. Rachidi, M. Bouzoubaa, L. El Mortaji, B. Boussouab and A. Bensaid, Arabic user search query correction and expansion), in: *Proceedings of COPSTIC'03*, Rabat, December, 13–11, 2003.

[23] H. Rodroguez, D. Farwell, J. Farreres, M. Bertran, M. Alkhalifa, M. Marta, M. Antonia, W. Black, S. Elkateb, J. Kirk, A. Pease, P. Vossen and C. Fellbaum, Arabic WordNet: current state and future extensions, *Proceedings of The Fourth Global WordNet Conference*, Szeged, Hungary, 2008.

[24] K. Shaalan, S. Al-Sheikh and F. Oroumchian, Query expansion based-on similarity of terms for improving Arabic information retrieval, in: Z. Shi, D. Leake and S. Vadera (eds.), *Intelligent Information Processing VI: Proceedings of 7th IFIP TC12 International Conference*, pp. 167–176, Springer, Heidelberg, 2012.

[25] M. Steyvers and J. Tenenbaum, The large-scale structures of semantic networks: statistical analysis and a model of semantic growth, *Cognit. Sci.* **25** (2005), 78–41.

[26] M. Symonds, P. Bruza, D. L. Sitbon and I. Turner, Tensor query expansion: a cognitively motivated relevance model, in: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM*, ACM, New York, NY, USA, 2012.