REPRESENTATION OF ARABIC WORDS
An Approach Towards Probabilistic Root-Pattern Relationships

Bassam Haddad
Faculty of Information Technology, Department of Computer Science, University of Petra, P.O.BOX 3034, Amman, Jordan

Keywords: Arabic NLP, Morphological Analysis, Root-Pattern Analysis, Statistical Language Model.

Abstract: In the traditional Arabic NLP a root-pattern relationship has generally been considered as a simple relationship, whereas the potential aspect of considering it as a statistical measure has extensively been neglected and even never formally considered. This paper attempts therefore to explore some issues involved in considering the classical phenomenon of Arabic root-pattern relationships as probabilistic measures. Some novel probabilistic measures in the context of Arabic NLP will be introduced with respect of their semantic potential as uncertain relations capturing some root related Arabic word-forms probabilistically.

1 INTRODUCTION

Arabic morphology corresponds to a singular class of morphological systems. It exhibits clear *non-concatenative* features; whereas manipulating the root letters is decisive for forming the majority of Arabic words. Roots represent the highest level of abstraction for a word basic meaning. Words can be morphologically classified into three classes of lexical words: *Basic Derivative*, *Rigid* (*Non-Derivative) and Arabized Arabic Words* (Haddad B., 2007).

Basic Derivative Arabic Words form the overwhelming majority of the Arabic lexical vocabulary. Most of these words can be generated from a *templatic tri-literal root* (/فعل/, f‘l) or the *quadri-literal root* (/فعلل/, f‘ll) by adding *consistent prefixes* and *suffixes* or *filling vowels* in a predetermined *pattern form*. *Non-Derivative* words include the lexical non-inflectional word types such as pronouns, adverbs, particles besides *stem words*, which cannot be reduced to a known root, whereas *Arabized Basic Words* consist of words without Arabic origin such as (/انترنت/, Internet). *Arabized Basic Words*, and *Non-Derivative Words*, do not linguistically exhibit a clear root-pattern relationship. This paper will focus the research interest on *Basic Derivative Words* and their semantic potential towards building novel probabilistic measures providing Arabic NLP with *statistical measures*. Furthermore, this paper will attempt to present formal description of these measures in the context of their applications in Arabic NLP, such as supporting morphological analysis, word-sense disambiguation, Non-Word detection and correction, information retrieval and others.

1.1 Related Research

The status of research on computational Arabic is limited compared to European languages, which have benefited from the broad research in this field. For the last decades, concentration on Arabic Language Processing has been focused on the *symbolic methods*, whereas the most effort has been focused extensively on *morphological analysis* (Beesley, 2001; Dichy and Farghaly, 2003; and many others), moderate on *syntax* (Ditters, 2001; Shaalan, 2005; and others) and relatively poor on *semantic* (Haddad B., 2007).

In the meanwhile, there are some attempts devoted to statistical methods utilizing traditional stochastic language models such as *HMM*, *Bayes Theorem* and *N-Gram Analysis* in word-sense *disambiguation*, *Arabic diacritization*, *Part of speech tagging* (Yaseen et Al., 2006), *Machine translation* (Shafer and Yarowsky, 2003) and others.

This paper is proceeding from the concept of *root-pattern analysis* as characteristic feature for

Haddad B. (2009).
REPRESENTATION OF ARABIC WORDS - An Approach Towards Probabilistic Root-Pattern Relationships.
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development*, pages 147-152
DOI: 10.5220/0002287101470152

representing the *majority of Arabic words*, whereas the major research contribution of this paper lies in extending the classical view of such aspect from simple lexical root-pattern relationship to *binary uncertain rule expressing predictive values* based on analysis of *frequency of occurrence*. In this context these root-pattern and pattern-root *probabilistic relations* correspond to the *point-valued binary fuzzy relation representing associative medical relationships* (Haddad, 2002).

Furthermore, this paper proposes to rewrite some *controversial basic concepts*, form procedural or functional point of view; whereas it is to hope that such formal concepts might serve as a possible source for finding f*ormal* standard descriptions of the *divisive notations* and descriptions found in literature of the Arabic computational community.

2 GENERATING BASIC WORDS

In this paper the focus of attention is the *class of derivative words*, which represent the major class of the Arabic word system. In the following some preliminary and basic notation are introduced:

The set of all Arabic roots and patterns will be represented by $\mathcal{R}$ and $\mathcal{P}_\mathcal{T}$ receptively :

$$\mathcal{R} = \{r_1, r_2, r_3, \ldots, r_r\}, \quad \mathcal{PT} = \{pt_1, pt_2, pt_3, \ldots, pt_{pt}\} \tag{1}$$

Let furthermore Θ_{root} be *a root substitution* replacing the root letters with letters occurring in a pattern.

Definition 1 (Templatic Root-Pattern Substitution).

Let (ف$_r$, f_r), (ع$_r$, ‘$_r$) and (ل$_r$, l_r) *be the templatic* root literals and $pt_i \in \mathcal{P}_\mathcal{T}$ *containing the templatic root literals, then a templatic root-pattern substitution is defined as*

$$\Theta_{root} = \{(ف_r, f_r)/(ف_{pt}, f_{pt}), (ع_r, ‘_r)/(ع_{pt}, ‘_{pt}), (ل_r, l_r)/(ل_{pt}, l_{pt}) \} \tag{2}$$

(See transcription in Appendix A)

Definition 2 (Instantiation a Templatic Root-Pattern Relationship).

Let $r_i \in \mathcal{R}$ and $pt_j \in \mathcal{PT}$ *then a basic word can be generated by application a templatic root-pattern substitution* "$\Theta root$ " *to the pattern* pt_j*:*

$$pt_j\, \Theta_{root}\, (ri) \tag{3}$$

Most Arabic words can be generated from the templatic *tri-literal root* (/فعل/, f‘l) or the *quadri-literal root* (/فعلل/, f‘ll). On the other hand, for each valid Arabic root, r_i , there is a *certain number of consistent patterns*, pt_j, with which a root can be instantiated. Therefore, a lexical derivative Arabic word can be understood as a result of *applying* a *substitution* of a root literal with the corresponding consistent pattern literals. Such a substitution can be regarded as a *transformation operation* of a root into a pattern word or an instantiation of template with root letters.

Example:
Let (/فاعل/, fā‘il) , (/مَفعول/, maf‘ūl) be patterns $\in \mathcal{P}_\mathcal{T}$ and (/كَتَب/, ktb, Writing) $\in \mathcal{R}$, then the application of the templatic root substitution to the patterns (/فاعل/, fā‘il) and (/مَفعول/, maf‘ūl) generates the following words respectively:

(/فاعِل/, **fā‘il**) Θ_{root} (/كَتَب/, **ktb,** Writing) =

(/كاتب/, **kātib,** Writer).

(/مَفعول/, **maf‘ūl**)Θ_{root} (/كَتَب/, ktb, Writing) =

(/ مكتوب /, **maktūb,** Letter).

The generated words are still basic words and represent basic stem words without considering morpho-syntactic and morphogramephic rules such as defection rules. A derivative Arabic word can in general be considered as an *incremental application* of different level of such rules to a root such Phonetic, N-Gram, Morph-Syntactic rules.

Introducing *further formal details* for these aspects such as the applicative feature of generating words based on root-pattern substitutions considering phonetic Morho-Syntactic rules *exceeds scope of this paper*. The focus of attention of this presentation is centered on *simple basic derivative words* in the context of establishing root- pattern relationships.

2.1 Representing Words as Root-Pattern Relations

In the traditional Arabic computational NLP community, root-pattern relationship is generally considered *from lexical look-up point of view*; i.e. a binary relationship expressing simply the presence of a root with a pattern or not. However, as it is well-known; due to *historical* reasons and difficulties of presenting *short vowels* without *diacritics*, the overwhelming written Arabic texts are *not vocalized*.

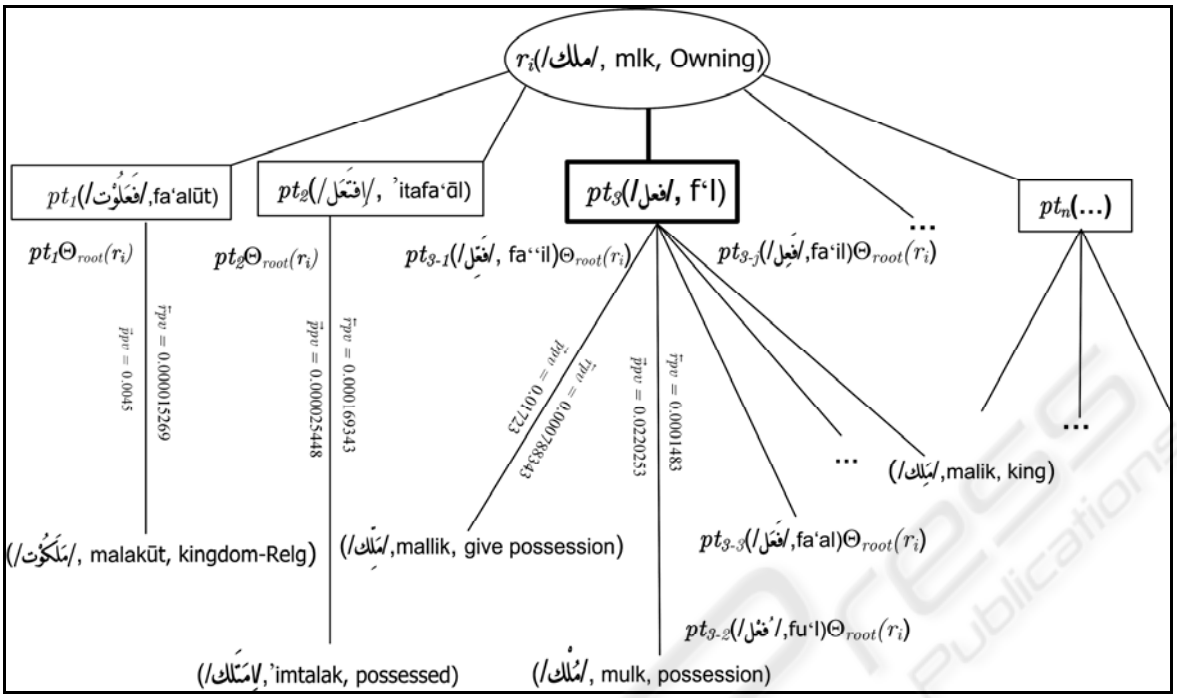

Figure 1: Root-Pattern Instantiations as Applicative Function Representation for the three radical root ri(/ملك/,mlk, Owning). Some root and pattern predicative values $\vec{r}pv$, $\vec{p}pv$ are depicted.

Considering additionally the fact that an Arabic root usually occurs with many different not vocalized patterns, would *exemplify* the main reason for the *strong ambiguity* and in particular in the lexical level in Arabic.

Such *ambiguity* is two fold in the sense, that for *one root* there will be many possible *unvocalized patterns,* and for a *one pattern* there might be more than one *possible root* whereas each root-pattern relationship might represent different word senses. The first type represents some kind of *polysemy.* In Figure1, the pattern pt_3 (/فعل/, f'l); due to the missing diacritics or short vocalizations on lexical level, is ambiguous and it can be interpreted in different ways. For example, in the above figure, application of the root r_i=(/ملك/, mlk, Owning) to different possible patterns might produce many different *instantiations* for root such as

$pt_{3\text{-}1}$(/فَعّل/, fa''il)$\Theta_{root}(r_i)$=(/مَلّك/,mallik, give possession).

$pt_{3\text{-}2}$(/فُعْل/,fu'l)$\Theta_{root}(r_i)$ =(/مُلْك/, mulk, possession).

$pt_{3\text{-}3}$(/فَعَل/,fa'al)$\Theta_{root}(r_i)$= (/ مَلَك/, malak, angel).

$pt_{3\text{-}j}$(/فِعِل/,fa'il) $\Theta_{root}(r_i)$= (/مَلِك/, malik, king)

and many other possibilities

Resolving such ambiguities based on *semantic or selection restrictions* and *dictionary look-up* is complex and needs in may cases exhaustive search.

In his approach, this paper is proposing to extend the representation of such relationships using novel *probabilistic root-pattern relations*, considering the possibility of *extending* this model to work on the *discourse representation* level within a N-gram analysis towards a hybrid approach.

3 REPRESENTING WORDS AS PROBABLISITC RELATIONS

As *patterns* or *templates* are significant for generating correct derivative words, *root-pattern and pattern-root relationships* in form of compatible or consistent rules can be established. Based on *frequency of occurrence* of a root with a pattern and occurrence of a pattern with a specific root, a probabilistic root-pattern and pattern-root relationship can be represented.

Definition 3 (Pattern-Predictive and Root-Predictive Values).

Let $r_i \in \mathcal{R}$, $pt_j \in \mathcal{PT}$ *then Root-Pattern Relationships*

can be established as follows:

$$\overrightarrow{\text{Root}}_{\text{PT}} \triangleq \{((r_i, pt_j), \vec{p}pv_{ij}) \mid (r_i, pt_j) \in \mathcal{R} \times \mathcal{PT}\} \quad (4)$$

where $\vec{p}pv_{i\text{-}j} \triangleq \text{P}(pt_j / r_i)$

$$\overleftarrow{\text{PT}}_{\text{Root}} \triangleq \{((r_i, pt_j), \overleftarrow{r}pv_{ij}) \mid (r_i, pt_j) \in \mathcal{R} \times \mathcal{PT}\} \quad (5)$$

where $\overleftarrow{r}pv_{ij} \triangleq \text{P}(r_i / pt_j)$

$\overrightarrow{\text{Root}}_{\text{PT}}$ can be interpreted as *uncertain forward binary rules* where as $\overleftarrow{\text{PT}}_{\text{Root}}$ can be interpreted as *uncertain backwards binary* rules.

Example:

Let r_i=(/كتب/, ktb, Writing)$\in\mathcal{R}$, pt_j=(/مَفْعُوْل/, mafūʻl)$\in\mathcal{PT}$ then based on $\text{P}(pt_j / r_i)$ we can establish a *binary uncertain relation* expressing the probability for predicting the instantiation of the pattern pt_j =(/مَفْعُوْل/, mafūʻl) with the given root such as

$$(/\text{كتب}/, \text{ktb}, \text{Writing}) \xrightarrow[ppv_{ij}]{} (/\text{مَفْعُوْل}/, \text{mafūʻl})$$

On the other hand, we can establish a binary uncertain relation expressing the *probability* for predicting that the *instantiated root* in the pattern (/مَفْعُوْل/, mafūʻl), is the root (/كتب/, ktb, Writing):

$$(/\text{كتب}/, \text{ktb}, \text{Writing}) \xleftarrow{rpv_{ij}} (/\text{مَفْعُوْل}/, \text{mafūʻl})$$

Table 1: Samples of some computed root predictive values, $\overleftarrow{r}pv$, and pattern predictive values, $\vec{p}pv$, based on a templaic root-pattern substitution for the root (/كتب/).

j	*ptj*	$\overleftarrow{r}pv_{(\text{بتك})_j}$	$\vec{p}pv_{(\text{بتك})_j}$
1	(/فَاعَلُ/, fāʻalu)	0.00055428	0.00003988
2	(/فَاعَلٌ/, fāʻalun)	0.00175608	0.00004226
3	(/فَاعَلَ/, fāʻala)	0.00012753	0.00001161
4	(/فَعَالٌ/, faʻālun)	0.00237203	0.00308527
5	(/فَعَالِ/, faʻālin)	0.00504853	0.00336621
6	(/مِفْعَلَةٌ/, mifʻalatun)	0.00244802	0.00003274
7	(/مِفْعَلَةً/, mifʻalatan)	0.00192429	0.00006547
8	(/مَفْعُوْل/, mafūʻl)	0.00524251	0.00140051
9	(/مَفْعُوْلٌ/, mafūʻlan)	0.01169770	0.00049225
10	(/فَعَالَةٌ/, faʻālati)	0.00071093	0.00001657

Based on the morphological analysis of a corpus containing 50544830 Arabic word-forms in one flat file about size 990 MB and Arabic dictionaries of about 31.5MB, normalized conditional probabilities have been assigned to 6860 Arabic roots in association with 650 patterns. The words have morphologically been pre-processed before computing the Root and Pattern Predictive Values. The data has been analyzed by ATW morphological analyzer, (http://www.arabtext.ws/), whereas suffixes and prefixes, stems, patterns and roots were initially extracted to be a subject of the subsequentail statistical analysis.

3.1 Applications

The significance of the introduced values; i.e. *Root and Pattern Predictive Values* depends on the application being observed in solving some Arabic NLP problem. *Pattern Predictive Relations* might be interpreted as *forward uncertain rules*; and namely, as lexicon look-up is actually a *root-based* search process, due to historical and lexicographical organizational reasons. On the other hand *Pattern Predictive Values* support processes involved in generating the most probable word patterns for some possible root; for example within a correcting process. This aspect can be significant for resolving some ambiguities and in ranking possible correcting candidates.

Root Predictive Values might come into effect in the case of generating the most probable roots, within a *root-extraction process* such as morphological analysis. These aspects might be extended to different possible applications such indexing, information retrieval and simple word-sense disambiguation.

The author has already utilized the introduced predictive values in a *hybrid approach* to detect and correct *non-words* in Arabic. The results were very helpful in optimizing the root-extraction process and in particular if the words were strongly deformed; whereas *Pattern Predictive Values* were significant in *ranking* and generating the most probable word candidates as possible correction. One of the most *interesting outcomes* of integrating these measures within this project were the quality of the results; as they have supported producing *accurate and natural* correcting candidates compared to standard spell-checkers such as Arabic MS-Word; details are found in (Haddad B. and Yaseen M, 2007).

4 CONCLUSIONS

This paper is an attempt to provide Arabic NLP with new probabilistic measures supporting issues involved in generating the most probable word patterns and roots on the lexical level. Based on statistical analysis of morphologically pre-processed corpus containing 50544830 Arabic word-forms, root and pattern predictive values have been estimated and assigned to 6860 Arabic roots in association with 650 patterns. In this context *Root and Pattern Predictive Values were introduced,* which might be interpreted as uncertain binary relations. The applicability of theses measures is wide-ranging such as supporting morphological analysis, word-sense disambiguation and non-word detection and correction on the lexical level, whereas syntactical cases of the pattern can also be considered.

These values have successfully been utilized in a hybrid approach to detect and correct Arabic Non-Words.

One interesting aspect of introducing these measures lies in the fact that root-pattern phenomenon of Arabic has directly been considered within a statistical model, which might reflect more natural result than pure and general purpose N-Gram analysis, used by different Arabic researchers.

On the other side, despite the fact that this model has considered the isolated morpho-syntactical pattern forms and their expected roots, it needs to be integrated within a discourse representational statistical language model to support more context depended applications such as deep semantic analysis and others. Presenting a *comprehensive model* based on the introduced measures exceeds the scope of this paper. The author is working on pursuing this objective considering more aspects, which can benefit from the presented measures besides investigating additional measures based on exploring the semantic potential of the introduced measures statistically.

REFERENCES

Beesley K. B, 2001. Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans 2001. In *ACL/EACL01, Conference of the European Chapter, Workshop: Arabic Language Processing: Status and Prospect.* France, Morgan Kaufman Publisher 2001.

Dichy Joseph and Farghaly A., 2003. Roots vs. Stems plus Grammar-Lexis Specifications: on what basis should a multilingual lexical databases centred on Arabic be built? In *MT SUMMIT IX, Workshop on Machine Translation for Semitic Languages: Issues and Approaches.* New Orleans, USA 2003, AMTA.

Ditters E., 2001. A Formal Grammar for the Description of Sentences Structures in Modern Standard Arabic. In *ACL/EACL01, Conference of the European Chapter, Workshop: Arabic Language Processing: Status and Prospect.* France, Morgan Kaufman Publisher 2001.

Fischer W., 19972. *Grammatik des Klassischen Arabisch*. Otto Harrassowitz, Wiesbaden.

Haddad B., 2007. Semantic Representation of Arabic: A logical Approach towards Compositionality and Generalized Arabic Quantifiers. In *International Journal of computer processing of oriental languages, IJCPOL 20(1) 2007***.** World Scientific Publishing.

Haddad Bassam and Yaseen M., 2007. Detection and Correction of Non-Words in Arabic: A Hybrid Approach. In *International Journal of Computer Processing of Oriental Languages, IJCPOL, Vol. 20, Number 4, December 2007*. World Scientific Publishing.

Haddad B., 2002. Representing the Ignorance about the Uncertainty in Associative Medical Relationships: An IBFR Approach. In *The 6th World MultiConference on Systemics*. IIIS, SCI 2002, Florida, USA 2002.

Shaalan K., 2005. GramChek: a grammar checker for Arabic. In *Software Practice and Experience*. John Wiley & sons Ltd., UK, 35(7):643-665, June 2005.

Shafer Charles and Yarowsky David, 2003. A Two-Level Syntax-Based Approach to Arabic-English Statistical Machine Translation. In *MT SUMMIT IX, Workshop on Machine Translation for Semitic Languages: Issues and Approaches,* new Orleans, USA 2003, AMTA.

Yaseen M., Atiyya M., Bendahman C., Maegaard B., Choukri K., Paulsson N., Haamid S., Fersøe H., Krauwer S., Rashwan M., Haddad B., Mukbel C.,Mouradi A., Ali A., Shahin M., Ragheb A., 2006. Building Annotated Written and Spoken Arabic Corpora Resources in NEMLAR Project. In *The Fifth International Conference on Language Resources and Evaluation.* LREC-2006, Genoa-Italy.

APPENDIX A

Transcription of Arabic Letters based on DIN and (Fischer, 1972). Long vowels are represented through the letters (ا, ā), (ى, ī) and (و, ū), while short vowels as follows: (fatḥa, ◌َ, a), (kasrah, ◌ِ, i) and (ḍammah, ◌ُ, u).

Letter	Transcription	Name
ء	‘	hamza
ا	Ā	alif’
ب	B	’bā

Letter	Transcription	Name
ت	T	‘tā
ث	ṯ	ṯā’
ح	ḥ	ḥā’
خ	ḫ	ḫā’
د	d	dāl
ذ	ḏ	ḏāl
ر	r	rā’
ز	z	zāy
س	s	sān
ش	š	šīn
ص	ṣ	ṣād
ض	ḍ	ḍāḍ
ط	ṭ	ṭ ā’
ظ	ẓ	ẓ ā’
ع	‘	‘ain
غ	ġ	ġain
ف	f	‘f ā
ق	q	qāf
ك	k	kāf
ل	l	lām
م	m	mīm
ن	n	nūn
ه	h	’hā
و	w, ū	wāw
ى	y, ī	’y ā