Assessment Criteria for Benchmarking of Arabic Morphological Analyzers and Generators

Tarek Elghazaly and Abdelmawgoud M. Maabid

Department of Computer and Information Sciences,
Institute of Statistical Studies and Research, Cairo University,
Egypt

tarek.elghazaly@cu.edu.eg

Abstract. Natural language processing applications are based on the morphology part. So they should meet some criteria in order to satisfy the required functionality. Assessing and evaluating of Arabic morphological systems depend on the input words and resulted output according to a predefined criteria to measure and analyze given system in order to study its weakness and strength, trying to find an Arabic morphological analyzer free from all mistakes. In this paper we developed the precise assessment criteria for Arabic morphological analyzers to be applied on a given Arabic morphological analyzers and stemming algorithms by voting, after running them on the sample documents selected as the gold standard.

Keywords: Morphology, Arabic morphology, NLP, morphology assessment criteria, stemmer, morphology benchmarking, analyzer, Arabic analyzer.

1 Introduction

Morphology in linguistics concerns with the study of the structure of words [1]. In other words, morphology is simply a term for that branch of linguistics concerned with the forms words take in their different uses and constructions [2].

Arabic is one of the languages having the characteristics that from one root the derivational and inflectional systems are able to produce a large number of words (lexical forms) each having specific patterns and semantics [3]. The root is a semantic abstraction consisting of two, three, or (less commonly) four consonants from which words are derived through the superimposition of templatic patterns [4]. Unfortunately if understanding is considered, un-diacritized words may make problems of meaning; where many words when they appears in un-diacritized text can have more than one meaning; these different meanings rises problems of ambiguity [5].

In Arabic, like other Semitic languages, word surface forms may include affixes, concatenated to inflected stems. In nouns, prefixes include conjunctions ("و" "and", فـ "and, so"), prepositions ("بـ" "by, with", "كـ" "like, such as", "لـ" "for, to") and a determiner, and suffixes include possessive pronouns. Verbal affixes include conjunction prefixes and negation, and suffixes include object pronouns. Either object or possessive pronouns can be captured by an indicator function for its presence or absence, as well as by the features that indicate their person, number and gender[6]. A large number of surface inflected forms can be generated by the combination of these features, making the morphological generation of these languages a non-trivial task [7].

Natural Languages processing and analysis improved substantially in recent years due to applying data intensive computational techniques [8]. However, state of the art approaches are essentially language specific stemmer (Morphology), considering every surface word in the language [9]. A shortcoming of this word-based analysis of the Arabic language is that it is sensitive to lack of data and information about Arabic words and it morphemes. This is an issue of importance as aligned corpora are an expensive resource, which is not abundantly available for many language analysis levels. This is particularly problematic for morphologically rich languages, where word stems are realized in many different surface forms, which exacerbates the hindering higher level of language analysis.

Morphological analysis can be performed by applying language specific rules. These may include a full-scale morphological analysis, or, when such resources are not available, simple heuristic rules, such as regarding the last few characters of a word as its morphological suffix. In this work, we will adapt some major assessment criteria for measuring advantage or drawback of any Arabic morphological system [10].

2 Background And Previous Work

We believe that this is the first proposed work to sum up assessment criteria for Arabic morphological analyzers and Generators. Several researches talked about building powerful stemmers for the Arabic language with accuracies normally exceeding 90% but none of these stemmers offer the source code and/or the datasets used. It is therefore difficult to verify such claims or make a comparison between different stemmers without having the full description of the proposed method or the source code for the implementation of the algorithm [11]. In this section we review some efforts in this direction.

Mohammed N. Al-Kabi and Qasem A. Al-Radaideh [11] proposed analysis of the accuracy and strength of four stemmers for the Arabic language using one metric for accuracy and four other metrics for strength as following:

- The first metric called empirical evaluation (EE), which represents a percentage of the correct roots produced by the stemmer under consideration.
- The mean number of words per conflation class (MWC) depends on the number of words processed.
- Index compression factor (ICF) represents the extent to which a collection of unique words is reduced (compressed) by stemming.
- Word change factor (WCF) represents the proportion of the words in a sample that have been changed in any way by the stemming process.

– The mean number of characters removed in forming stems (Average CR): Usually strong stemmers remove more characters from words to form stems.

Azze Al-din Al-Mazroui, et al. [12] proposed a specification of morphological analysis system in the Arabic language. In this study the researcher outlined the general characteristic that has to consider during process and building Arabic morphological system in terms of input, analysis and output. The study doesn't provide any criteria or automation to compare different systems.

Dassouki [13] proposed a tabulate items as mechanism for assessing morphological analyzer in terms of development of the system speed, input, output, integrating with other applications and capabilities of analyzing new and non-Arabic words. The study doesn't provide any criteria for these selected terms.

William B. Frakes and Christopher J. Fox [14] evaluated the strength and similarity among, four affix removal stemming algorithms. Strength and similarity were evaluated in different ways, including new metrics based on the Hamming distance measure. Data was collected on stemmer outputs for a list of 49,656 English words derived from the UNIX spelling dictionary and the Moby corpus. The study doesn't provide any criteria for these selected measures and it is specific to English stemmers.

3 Proposed Assessment Criteria of Arabic Morphological Systems

Assessing and evaluating Arabic morphological systems depends on the *input* words and resulted *output* [12] according to a predefined criteria to measure and analyze given system in order to study its weakness and strength, trying to find an Arabic morphological analyzer free from all mistakes. Then we will apply these criteria on some of existing available systems; these criticisms will not detract from its value and effectiveness.

3.1 Input

A very fundamental problem with software testing is that testing under all combinations of inputs and preconditions (initial state) is not feasible, even with a simple product. The input can be considered as bulk of text passed to the system in form of word or phrase fully or partially diacritized.

The possibility of analyzing the modern standard texts

Most western scholars distinguish two standard varieties of the Arabic language: the Classical Arabic (CA) of the Qur'an and early Islamic (7th to 9th centuries) literature, and Modern Standard Arabic (MSA), the standard language in use today [15]. The modern standard language is based on the Classical language. Most Arabs consider the two varieties to be two registers of one language, although the two registers can be described in Arabic as (MSA) and (CA) [16].

The possibility of analyzing the common error words

Common typing errors "common error words" are those words mistyped but are traditionally considered correct; typically a feminine ending character “ـة” written without dots “ـه”, the dotless “ى” instead of “ي” and the letter “ا”without hamza instead of "أ"; for e example word "احمد" can be read and understood correctly as "أحمد" while the first one is linguistically mistyped [17].

The possibility of analyzing new words (Neologisms)

Neologisms are often created by combining existing words or by giving words new and unique suffixes or prefixes. Portmanteaux "حقائب السفر" are combined words that are sometimes used commonly. Neologisms also can be created through abbreviation or acronym, by intentionally rhyming with existing words or simply through playing with sounds.

Neologisms can become popular through memetics, by way of mass media, the Internet, and word of mouth, including academic discourse in many fields renowned for their use of distinctive jargon, and often become accepted parts of the language. Other times, however, they disappear from common use just as readily as they appeared. Whether a neologism continues as part of the language depends on many factors, probably the most important of which is acceptance by the public. It is unusual, however, for a word to enter common use if it does not resemble another word or words in an identifiable way.

When a word or phrase is no longer "new", it is no longer a neologism. Neologisms may take decades to become "old", however. Opinions differ on exactly how old a word must be to cease being considered a neologism.

Neologisms analysis in morphological system measures the capability of processing the new Arabic words which can be added later to morphological systems’ predefined knowledge base.

Processing of Arabized and transliterated words

Transliteration is a subset of hermeneutics. It is a form of translation, and is the practice of converting a text from one script into another. From an information-theoretical point of view, systematic transliteration is a mapping from one system of writing into another, word by word, or ideally letter by letter. Transliteration attempts to use a one-to-one correspondence and be exact, so that an informed reader should be able to reconstruct the original spelling of unknown transliterated words. Ideally, reverse transliteration is possible.

Transliteration is opposed to transcription, which specifically maps the sounds of one language to the best matching script of another language. Still, most systems of transliteration map the letters of the source script to letters pronounced similarly in the goal script, for some specific pair of source and goal language. If the relations between letters and sounds are similar in both languages, a transliteration may be (almost) the same as a transcription. In practice, there are also some mixed transliteration/transcription systems that transliterate a part of the original script and transcribe the rest [13].

In Arabic transliteration is writing non-Arabic words by Arabic alphabet characters as ‘فاكس’ “Fax” in English and “انترنت” “Internet” In English.

Processing of non-tripartite verbs

Arabic verbs, as the verbs in other Semitic languages, are more complicated than those in most languages. A verb in Arabic is based on a set of three or four consonants called a root (trilateral or quadrilateral according to the number of consonants). The root communicates the basic meaning of the verb, e.g. "كتب" k-t-b "write", "قرأ" q-r-’ "read", and "أكل" ’-k-l "eat". Changes to the vowels in between the consonants, along with prefixes or suffixes, specify grammatical functions such as person, gender, number, tense, mood, and voice.

Arabic words are divided into three types: noun, verb, and particle. Nouns and verbs are derived from a closed set of around 10,000 roots. The roots are commonly three or four letters and are rarely five letters. Arabic nouns and verbs are derived from roots by applying templates to the roots to generate stems and then introducing prefixes and suffixes [6].

Assessing and evaluating Arabic considering the system capability of analyze quadrilateral and quinqueliteral verbs like “طمأن” "Reassure" and all possible cases of their forms of transitivity and weakness [12].

3.2 Output

Morphology output is all possible combination of affixes that produced a valid Arabic word, roots and patterns.

Covering analysis of all input words

- The system should cover all cases of analysis.
- Determine word types (pattern, root, stem and attached affixes) [12].
- Analyzing the words in all domains of the language (Geographic, Historical, Religion, and Math).
- Considering syntactic case of input word (within phrase)

Meet all possible cases for analysis

The system has to assume that the input word is a verb, name and character so it has to determine the followings:

- Verb: has to cover non- tripartite, quadrilateral, quinqueliteral with their forms of transitivity, augmentation, hollow…etc. [4].
- Name: has to cover names, infinitives, adjectives and adverbs.
- Particle: has to cover prepositions, conjunctions, vowel, and vocative particles.

Express grammatical function of the affixes

Affixes are those characters attached to the stem (prefix, suffix and infix) each has its own grammatical alternation of the stem attached.

Ambiguity and overlapping of syntactic cases

Many words in Arabic are homographic [5]: they have the same orthographic form, though the pronunciation is different. There are many recurrent factors that contributed to this problem. Among these factors are:

- Orthographic alternation operations (such as deletion and assimilation) frequently produce inflected forms that can belong to two or more different lemmas.
- Some lemmas are different only in that one of them has a doubled sound which is not explicit in writing. Arabic Form I and Form II are different only in that Form II has the middle sound doubled.
- Many inflectional operations underlie a slight change in pronunciation without any explicit orthographical effect due to lack of short vowels (diacritics).
- Some prefixes and suffixes can be homographic with each other. The prefix t can indicate 3rd person feminine or 2nd person masculine.
- Prefixes and suffixes can accidentally produce a form that is homographic with another full form word. This is termed “coincidental identity”
- Similarly, clitics can accidentally produce a form that is homographic with another full word.
- There are also the usual homographs of uninflected words with/without the same pronunciation, which have different meanings and usually different POS’s.

That means determining the lack of morphological knowledge of the word analyst; in case of partially diacritized or non-diacritized words, the ambiguity problem may appear, so, the better is to determine all possible cases of the input word; as an example the work “رب” many be either “رَبٌّ” (God) or “رُبَّ” (maybe).

Identifying the root of the word and determining all possible roots for the analyzed word

Right root identification of the input word, and with all generated words the system has to be capable to determine their roots and patterns.

Grammatical errors and misspellings in the context of the expression of results of the analysis

The output representation of the system has to be error free in terms of expression and representation of output result.

Cover all possible cases of syntactic word analyst

The system also should be represent and explain the analysis result of each of analyzed word and there generated words.

Consistency between analyzed word and its patterns

The system should produce correct and consistent patterns for the analyzed and generated words.

The result has to be coming from Arabic dictionary
The system should combine the Arabic morphological rules while processing the word with its knowledgebase to reflect a better analysis and generation which measures the trust of morphological analysis result.

3.3 System Architecture and Design

Percentage of non-reliance on predefined knowledgebase of affixes, roots and patterns
An affix is a morpheme that is attached to a word stem to form a new word. Affixes may be derivational, like English -ness and pre-, or inflectional, like English plural -s and past tense -ed. They are bound morphemes by definition; prefixes and suffixes may be separable affixes. Affixation is, thus, the linguistic process speakers use to form different words by adding morphemes (affixation) at the beginning (prefixiation), the middle (infixation) or the end (suffix) of words.

Percentage of non-reliance on common words (Stop List)
In Natural Language Processing (NLP), stop words are words which are filtered out prior to, or after, processing of natural language data. Any group of words can be chosen as the stop words for a given purpose. Common words (stop word) are the words that are frequently used in Arabic text with the same meaning such as day names, month names, numbers names, adverbs… etc.

Processing Speed
In software engineering, performance testing is in general testing performed to determine how a system performs in terms of responsiveness and stability under a particular workload. It can also serve to measure, investigate, validate or verify other quality attributes of the system, such as scalability, reliability and resource usage.

Performance testing is a subset of performance engineering, an emerging computer science practice which strives to build performance into the implementation, design and architecture of a system.

The processing speed can be measured by how many words processed per second.

Ease of use and integration with larger applications
In engineering, system integration is the bringing together of the component subsystems into one system and ensuring that the subsystems function together as a system. In information technology, systems integration is the process of linking together different computing systems and software applications physically or functionally, to act as a coordinated whole.

- How much the system is capable for use and what are the prerequisites for the system to run.
- The ability to integrate the system within larger applications.

- The ability of modifying some of the system behavior of output or even input procedures and functions. (Customization).
- The ability to add inputs to the system knowledgebase.

Availability and documentation

Software documentation or source code documentation is written text that accompanies computer software. It either explains how it operates or how to use it, or may mean different things to people in different roles.

In terms of Arabic morphological system, it measures the availability of the system and its algorithms for newcomer and researchers considering the cost of commercial systems.

User interface (English-Arabic)

The user interface, in the industrial design field of human–machine interaction, is the space where interaction between humans and machines occurs. The goal of interaction between a human and a machine at the user interface is effective operation and control of the machine, and feedback from the machine which aids the operator in making operational decisions. User interfaces exist for various systems, and provide a means of:

- Input, allowing the users to manipulate a system
- Output, allowing the system to indicate the effects of the users' manipulation

Generally, the goal of human-machine interaction engineering is to produce a user interface which makes it easy, efficient, and enjoyable to operate a machine in the way which produces the desired result. This generally means that the operator needs to provide minimal input to achieve the desired output, and also that the machine minimizes undesired outputs to the human.

There are two major factors for judging morphological system interface as follows:

- The Interface language of system itself.
- The language used to represent the output of the system in case of analysis or generation.

Encoding and word representation

Identifying the character encoding used in the system itself for processing and representing the data. As Arabic letters need to be represented in Unicode set; some systems need to transliterate the input as a preparation for processing step and then revert the transliterated results into Arabic to match user input and user interface.

4 Application of the Proposed Assessment Criteria

Assessments are carried out by executing some of the available Arabic morphological analyzers on a randomly selected Arabic political news article, an Arabic Sport News article “from Al-Ahram newsletter” and the Chapter number 36 of the Holy Qur’an “سورة يس Surah Yassin” with total of 11000 distinct words. We then manually extracted the roots of the test documents’ words to compare results from different analyzers, thus creating our baseline test set. Roots extracted were then checked manually in an Arabic dictionary. Voting weights are assigned to each assessment item (assigned Score) in order to accurately make comparisons between these algorithms. Each assessment item has to be applied and calculated as per the result of applying the analysis to the sample input words. Table 1, shows assessment items where the voting mark of each individual item is assigned score of 100points. Here is the step by step procedure of executing the assessment criteria:

1. Manually extract the roots of the test documents’ words.
2. Assign voting mark for each assessment item.
3. Manually check the extracted roots against Arabic dictionary.
4. Apply each assessment item separately on each of Arabic Morphological Analyzer.
5. For the output results, check them manually against Arabic dictionary.

Finally, the assessment factors can be separately applied on each of Arabic Morphological Analyzer where all factors can be assigned score with a maximum value of 100 marks. Each assessment factor will be applied and calculated as per Analyzer result of applying the analysis of the sample document words.

Table 1. Assigned scores of the assessment factors.

Cat.	No.	Assessment Criteria	Score %
Input	1	The possibility of analyzing the standard and modern texts	100
	2	The possibility of analyzing the common error words	100
	3	The possibility of analyzing new words	100
	4	Processing of Arabized and transliterated words	100
	5	Processing of non- tripartite verbs.	100
Output	6	Covering analysis of all input words	100
	7	Meet all possible cases for analysis	100
	8	Express grammatical function of the affixes	100
	9	Ambiguity and Overlapping of syntactic cases	100
	10	Identifying the root of the word and determining all possible roots	100
	11	Grammatical errors and misspellings in the context of the results of the analysis	100
	12	Cover all possible cases of syntactic word analyst	100
	13	Consistency between analyzed word and its patterns	100
	14	The result has to be coming from Arabic dictionary	100
System Architectu	15	Percentage of non-reliance on predefined knowledgebase of affixes	100
	16	Percentage of non-reliance on common words	100
	17	Processing Speed	100

Cat.	No.	Assessment Criteria	Score %
	18	Ease of use and integration with larger applications	100
	19	Availability, documentation and customization	100
	20	User Interface (English - Arabic)	100
	21	Encoding and word representation	100
Sum			2200

5 Experiments and Results

Experiments are done by executing some of existing and available Arabic morphological systems on a randomly selected contemporary Arabic political news article, Arabic Sport News article "from Al-Ahram newsletter" and the first 15 verses of chapter number 36 of the Holy Qur'an "Souraht Yassin". Each test document contains domain specific words and represents contemporary and standard Arabic. The test documents contain 540 distinct token. We manually extracted the roots of the test documents' words to compare results for each stemming algorithm. Roots extracted have been check against Arabic dictionary.

The analysis also show that function words such as "فى" "fi", "من" "min", "بين" "bian" are most frequent words in any Arabic text. In other hand, nonfunctional words with high frequency such as "الإفريقية" "al-afiriqiah", "القمة" "al-Qemah" and other words out of 30 most frequent tokens as shown in table I gives a general idea about the main topic of the article.

Simple tokenization is applied for the text of the gold standard documents can be used to test any algorithm smoothly and correctly.

Table 2. Assessment results.

Factor No.	Morphology System			
	Al-Khalil	***Sarf***	***AMA***	***Khoja***
1	75	NA	80	50
2	85	NA	90	20
3	30	NA	20	0
4	10	NA	5	0
5	90	NA	85	80
6	75	NA	80	70
7	87	NA	85	0
8	92	NA	80	0
9	90	NA	35	30
10	85	NA	95	30
11	85	NA	98	90
12	45	NA	40	0
13	80	NA	95	0
14	86	NA	97	80
15	0	0	0	0

16	0	0	0	0
17	35	0	0	30
18	60	60	30	60
19	70	85	0	70
20	50	50	50	50
21	50	50	10	10
Total	1280	245	1075	670

6 Conclusion and Future Research

The proposed assessment criteria are adapted to measure Arabic Morphological Analyzers with some features intended for integration with lager applications in natural language processing. Many other criteria can be added to the proposed items and may vary in weight and phase of testing; similar to the source code related metrics used for measuring the system as a product.

The stemming algorithms involved in the experiments agreed and generate analysis for simple roots that do not require detailed analysis. So, more detailed analysis and enhancements are recommended as future work.

Most stemming algorithms are designed for information retrieval systems where accuracy of the stemmers is not important issue [18]. On the other hand, accuracy is vital for natural language processing. The accuracy rates show that the best algorithm failed to achieve accuracy rate of more than 65%. This proves that more research is required.

References

1. Kiraz, G.A.: Computational Nonlinear Morphology with Emphasis on Semitic Languages. Studies in Natural Language Processing, ed. I. Branimir Boguraev, T.J. Watson Research Center and L.D.C. Steven Bird, University of Pennsylvania, The Edinburgh Building, Cambridge CB2 2RU, Cambridge, United Kingdom (2004)
2. Beesley, K.R.: Arabic Morphological Analysis on the Internet. In 6th International Conference and Exhibition on Multi-lingual Computing, Cambridge (1998)
3. Buckwalter, T.: Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium (2002)
4. Watson, J.C.E.: The Phonology and Morphology of Arabic. The phonology of the world's languages, ed. J. Durand, New York, United States: Oxford University Press (2007)
5. Mohammed, A.A.: An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks, School of Informatics, The University of Manchester (2006)
6. Darwish, K.: Building a Shallow Morphological Analyzer in One Day. In 40th Annual Meeting of the Association for Computational Linguistics (ACL-02), Philadelphia, PA, USA (2002)
7. Soudi, A., V. Cavalli-Sforza, and A. Jamari: A Computational Lexeme-Based Treatment of Arabic Morphology. In Arabic Natural Language Processing Workshop, Conference of the Association for Computational Linguistics (ACL 2001) Toulouse, France (2001)

8. Soudi, A., A.V.D. Bosch, and G.U. Neumann: Arabic Computational Morphology. Knowledge-based and Empirical Methods. Text, Speech and Language Technology, ed. N. Ide et al. Vol. 38, The Netherlands: Springer (2007)
9. Shaalan, K.F. and A.A. Rafea: Lexical Analysis of Inflected Arabic Words using Exhaustive Search of an Augmented Transition Network. Software Practice and Experience, Vol. 23(6) (1993)
10. Roark, B. and R. Sproat: Computational Approaches to Morphology and Syntax, United States: Oxford University Press, New York (2007)
11. Al-Kabi, M.N., Q.A. Al-Radaideh, and K.W. Akkawi: Benchmarking and assessing the performance of Arabic stemmers. Journal of Information Science, Vol. 37(111) (2011)
12. Mazrui, A., et al.: Morphological analysis system specifications. In Meeting of experts in computational morphological analyzers for the Arabic language, Damascus (2010)
13. Desouki, M.S.: Mechanism for assessing morphological analyzer. In Meeting of experts in computational morphological analyzers for the Arabic language, The Arab League Educational, Cultural and Scientific Organisation (ALECSO) - King Abdulaziz City for Science and Technology: Damascus (in Arabic) (2009)
14. Frakes, W.B. and C.J. Fox.: Strength and Similarity of Affix Removal Stemming Algorithms. In Proceedings of the Annual Conference on Research and Development in Information Retrieval, ACM SIGIR Forum (2003)
15. Mushira Eid, C.H.: Perspectives on Arabic Linguistics V: Papers from the Fifth Annual Symposium on Arabic Linguistics. Volume 5: John Benjamins Publishing Company (1993)
16. Elgibali, A., K. Versteegh, and M. Eid: Encyclopedia of Arabic Language and Linguistics. Brill Academic Pub. 3250 (2009)
17. Eid, M., V. Cantarino, and K. Walters: Perspectives on Arabic Linguistics VI: Papers from the Sixth Annual Symposium on Arabic Linguistics, Volume 4, John Benjamins Publishing Company, 238 (1994)
18. Sawalha, M. and E. Atwell.: Comparative Evaluation of Arabic Language Morphological Analysers and Stemmers. In COLING 2008 22nd International Conference on Comptational Linguistics, Manchester (2008)