Natural Language Engineering (2023), **29**, pp. 978–1003
doi:10.1017/S1351324922000444

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

A benchmark for evaluating Arabic word embedding models

Sane Yagi[1], Ashraf Elnagar[2,*] and Shehdeh Fareh[1]

[1]Department of Foreign Languages, University of Sharjah, Sharjah, UAE and [2]Department of Computer Science, University of Sharjah, Sharjah, UAE
*Corresponding author. E-mail: ashraf@sharjah.ac.ae

(Received 7 May 2021; revised 8 September 2022; accepted 12 September 2022; first published online 17 October 2022)

Abstract

Modelling the distributional semantics of such a morphologically rich language as Arabic needs to take into account its introflexive, fusional, and inflectional nature attributes that make up its combinatorial sequences and substitutional paradigms. To evaluate such word distributional models, the benchmarks that have been used thus far in Arabic have mimicked those in English. This paper reports on a benchmark that we designed to reflect linguistic patterns in both Contemporary Arabic and Classical Arabic, the first being a cover term for written and spoken Modern Standard Arabic, while the second for pre-modern Arabic. The analogy items we included in this benchmark are chosen in a transparent manner such that they would capture the major features of nouns and verbs; derivational and inflectional morphology; high-, middle-, and low-frequency patterns and lexical items; and morphosemantic, morphosyntactic, and semantic dimensions of the language. All categories included in this benchmark are carefully selected to ensure proper representation of the language. The benchmark consists of 45 roots of the trilateral, all-consonantal, and semivowel-inclusive types; six morphosemantic patterns (ʾafʿala; iftaʿala; infaʿala; istafʿala; tafaʿʿala; and tafāʿala); five derivations (the verbal noun, active participle, and the contrasts in Masculine-Feminine; Feminine-Singular-Plural; Masculine-Singular-Plural); and morphosyntactic transformations (perfect and imperfect verbs conjugated for all pronouns); and lexical semantics (synonyms, antonyms, and hyponyms of nouns, verbs, and adjectives), as well as capital cities and currencies. All categories include an equal proportion of high-, medium-, and low-frequency items. For the purpose of validating the proposed benchmark, we developed a set of embedding models from different textual sources. Then, we tested them intrinsically using the proposed benchmark and extrinsically using two natural language processing tasks: Arabic Named Entity Recognition and Text Classification. The evaluation leads to the conclusion that the proposed benchmark is truly reflective of this morphologically rich language and discriminatory of word embeddings.

Keywords: Language Resources; Semantics; Similarity; Syntax; Arabic embedding models

1. Introduction

Distributional semantic representation of texts has become the norm in Natural Language Processing (NLP) since the publication of Mikolov *et al.* (2013). Word embeddings in the form of high-dimensional vectors in a semantic space are currently at the heart of most research in language processing. They are at the foundation of machine learning processing. Yet and despite their popularity, their evaluation metrics are questionable.

Not withstanding "problems with the appropriateness and informativeness of word analogy tests in current distributional word embedding evaluation" (Schluter 2018), analogical reasoning is "the second most popular method of word embeddings evaluation. . . [It] is based on the idea

that arithmetic operations in a word vector space could be predicted by humans" (Bakarov 2018). For analogical reasoning, scholars use the original Mikolov et al. evaluation dataset, named Google Analogy and Semantic-Syntactic Word Relationship Dataset.

We claim that an evaluation dataset that mirrors this Mikolov et al. dataset is inappropriate for Arabic as it fails to capitalize on the nature of this language. In this paper, we aim to substantiate our claim and propose a dataset for the evaluation of word embeddings that is more reflective of the nature of Arabic and therefore, more effective in evaluation.

Each word in the Arabic language belongs simultaneously to morphological, syntactic, and semantic paradigms that result from the word being derived from a root, coined by a morphological pattern, and associated with other words that form phrases and sentences. Current Arabic evaluation datasets are either near replicas or translations of Mikolov's dataset; they do not reflect the morphological, syntactic, and semantic paradigms that Arabic words participate in. It would not be feasible for such datasets to tell how good an Arabic word embedding is if they do not reflect its derivational and inflectional nature.

This paper aims to:

- show that the metrics currently used for the evaluation of Arabic embedding models are inadequate since this language has root-based, introflexive, fusional morphology, and inflectional syntax that they ignore.
- propose a modified benchmark that would account for the distinctive morphological and syntactic nature of the Arabic language.
- build 40 embedding models for a range of datasets using the algorithms of FastText (FT), Global Vectors (GloVe), and Word2Vec (W2V) in their Continuous Bag of Words (CBOW) and Skip-Gram (SG) architectures.
- demonstrate the reliability and effectiveness of this benchmark both intrinsically and extrinsically.

The remainder of this paper is organized as follows. The procedural definitions are presented in Section 2. In Section 3, we describe previous research on evaluation schemes of word embedding models. Next, we outline the components of the proposed benchmark. Then, we evaluate a variety of Arabic word embedding models intrinsically (using the proposed benchmark) and extrinsically on two NLP tasks in Section 5. Finally, we conclude with general remarks in Section 6.

2. Procedural definitions

We are focused in this paper on presenting a set of tests that will reflect linguistic patterns in both Contemporary Arabic and Classical Arabic. These tests will record the major features of nouns and verbs; derivational and inflectional morphology; high-, middle-, and low-frequency lexical patterns; as well as major morphosemantic, morphosyntactic, and semantic features of the language.

To achieve this goal, we may need to define a few technical terms that we will use in the process. We will define each briefly and illustrate it with examples that will ensure common understanding.

Modern Standard Arabic (MSA) is the modern form of literary Arabic that is associated with the Arab Renaissance which itself resulted from the cultural shock that the Napoleon invasion of Egypt in 1798 A.D. caused. It is the modern written form of Arabic, the medium of formal communication, and the lingua franca that is used in all Arab countries. It is the language of education that literary writers, journalists, public speakers, and religious leaders use at formal occasions.

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

Pre-Modern Arabic is any form of Arabic that was spoken or written prior to 1798 A.D., the date of Napoleon's invasion of Egypt.

Classical Arabic is the language of scholarship that was used from pre-Islamic times until 1798 A.D.

Contemporary Arabic is a term that is currently in circulation but we are using it here as a cover term for the written and spoken forms of Arabic. Contemporary Arabic for us includes all modern varieties of the language: MSA, Middle Arabic, and the dialects. It only excludes Pre-Modern Arabic.

Derivational morphology: The study of word formation as it relates to affixes that attach to a word to create a new word of a different meaning and/or word class. Affixes that create new words are called derivational affixes. For example, in English im-, -ful, -tion, and -ly in im-proper, skill-ful, construc-tion, and quick-ly are all derivational affixes. In Arabic, derivational morphology uses morphosemantic patterns to create new nouns and verbs from a root. For example, derivational morphology applies verb and noun morphosemantic patterns to the root *shrk* ش ر ك to produce, inter alia, the following words: أَفعَلَ 'af'ala, أَشرَكَ 'ashraka *'he got him to share'*; فاعَلَ 'fā'ala شارَكَ shāraka *'he shared with'*; تَفاعَلَ tafā'ala تَشارَكَ tashāraka *'together they went into partnership'*; فَعِيل 'fa'īl شَرِيك sharīk *'partner'*; فَعِلَة fa'ilah شَرِكَة sharikah *'company'*; مُفاعَلَة mufā'lah مُشارَكَة mushārakah *'partnership'*; فَعَل fa'al شَرَك sharak *'snare'*.

Morphosemantics is the interaction between morphology and semantics, between the form of a word and its meaning. Any change in the form of a word that results in change in meaning is morphosemantic. Therefore, interdigitating consonants and vowels between root radicals is a morphosemantic operation that results in significant changes in meaning, that is, all Arabic morphological patterns are the result of morphosemantic processes; hence, they are also called 'morphosemantic patterns'. For example, the root *K T B* ك ت ب *'to write'* may have the following morphosemantic forms: كاتِب kātib *'writer'*, مَكتُوب maktūb *'written'*, كِتابَة kitābah *'writing'*, مُكاتَبَة mukātabah *'correspondence'*, etc.

Morphosemantic template is also called morphosemantic pattern, morphological pattern, *al-wazn al-sarfi*. There are 10 verb patterns that are highly frequent in Arabic and about 60 noun patterns that are most productive. Here are some examples: أَفعَلَ 'af'ala '*caused an event*'; فاعَلَ fā'ala 'initiated an event'; تَفاعَلَ tafā'ala *'together participated in an event'*; فَعِيل fa'īl *'with an attribute of an event'*; فَعلَة fa'lah *'an instance of an event'*; مُفاعَلَة mufā'alah *'the act of participation in an event'*, etc.

Morphophonemic alteration (transformation): Changes in the form of a word because of a sequence of sounds that it contains, or changes in the pronunciation of a word because of a sequence of word parts. For example, the plural marker in English is pronounced as s, z, or əz depending on the final sound in the word. In Arabic, the definite article is pronounced as al or ass, arr, azz, att, etc. depending on the nature of the initial sound in the word that it attaches to.

Inflectional morphology: The study of word structure as it relates to affixes that do not change the word class or meaning. Such affixes are referred to as 'inflectional affixes'. In English, these are the -s at the end of the present tense verb when the subject is third person singular, the -ed at the end of a past tense verb, the -er at the end of a comparative adjective, etc. In Arabic verbs, inflectional affixes are those that indicate its tense, mode, subject, and object markers, and in Arabic nouns, they indicate definiteness, number, gender, and possession. For example, all affixes in this

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

word are inflections: تُشاطِرنَهُم tu-shāṭir-na-hum *imperfective-'share'-they (feminine, plural)-they (masculine, plural) 'they share with them'*.

Inflectional syntax: Sentence structure that relies more on inflectional affixes than on word order to express the grammatical roles of words.

Morphosyntax, on the other hand, is the interaction between morphology and syntax, between the form of a word and its grammatical role in a sentence. Morphosyntax is realized by the inflectional affixes, the affixes for definiteness, case, number, person, tense, and mood that are added to a stem to mark its grammatical attributes; that is, all word forms of a noun and all conjugations of a verb are the result of morphosyntactic operations since they are only inflected syntactically and without any alteration of meaning. In these words, for instance, al-**muhandis**-ūna 'the engineers', and yu-**jālis**-na-hum 'they socialize with them', all the prefixes and suffixes are added to the stems in bold by morphosyntactic rules. Unlike morphosemantic alteration, the affixes here do not alter the meanings of the words but adapt them to the contexts of use instead.

Introflexive language: A language that uses root and pattern to construct words by interdigitating vowels and/or consonants between the radicals of a root. Arabic is an introflexive language.

Fusional language: Arabic and Latin are fusional languages because they fuse such features as gender, case, and number; or tense, voice, mode, and pronominal subject in a manner which makes it difficult to segment a word into its constituent morphemes. For example, the prefix ya- and the suffix -u in ya-drus-u mark the mode of this verb as 'indicative', the tense as 'imperfect', the voice as 'active voice', and the subject as 'third person singular masculine'.

Active participle اسم الفاعِل ism al-fāʿil: A noun or adjective that denotes the doer of an action. It corresponds to English nouns that end with -er or -or and to adjectives that end with -ing. For example, عامِل āmil *'worker'*; مُدَرِّسُهُ mudarrisuhu *'teaching him'*; مُدرِك mudrik *'cognizant'*; مُرسِل mursil *'sender'*; مُراسِل murāsil *'correspondent'*, etc.

Verbal noun المَصدَر al-maṣdar: A morphological pattern that names the action of the verb it corresponds to and is usually translated in English as an infinitive. For example, صُعُوبَة ṣuʿūbah *'hardship'*; اِحتِواء iḥtiwā *'containment'*; اِكتِشاف iktishāf *'discovery'*; كِتابَة kitābah *'writing'*; مَعرِفَة maʿrifah *'knowledge'*; تَعزِيز taʿzīz *'reinforcement'*; دِفاع difāʿ *'defense'*; تَنافُس tanāfus *'rivalry'*; اِنقِياد inqiy ād *'docility'*; اِبتِداء ibtidāʿ *'initialization'*; اِستِثمار istithmār *'investment'*.

W-inclusive and y-inclusive roots: Roots that consist of a combination of consonants and semivowels (w, y). These are w-inclusive and y-inclusive roots: wy ي ع و *'to be cognizant'*; hwm م و ٥ *'to roam'*; d'w و ع د *'to call'*; yqn ن ق ي *'to give credence to'*; byt ت ي ب *'to dwell in'*; ḥny ي ن ح *'to bend'*.

All-consonantal imperfect verb: Present tense verb whose root consists of consonants only (صَحِيح سالِم ṣaḥīḥ sālim), for example: يَكتُبُ yaktubu *'he writes'*; تَلعَبُ talʿabu *'she plays'*; يَدرُسنَ yadrusna *'they (fem) study'*.

Clitic: A word part which is like a word in having a meaning of its own and like an affix in being always bound to another word. It may be a proclitic that attaches at the beginning of a word, and enclitic that attaches at the end of a word. For example, the Arabic definite article al- and the

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

conjunction wa- are proclitics, while the subject and object pronominal pronouns that attach at the end of a verb are enclitics.

Graphemic normalization: Representing all forms of a letter by one of them. For example, the hamza (أ ؤ ئ ء ر) is often normalized as (ء) and the alefs (ا ى) as (ا).

Gemination diacritic: The shadda symbol written as this superscript: ّ.

Vowel elongation: pronouncing the vowel longer than normal.

Metaphorical extension: Extending the meaning of a word by using it in a metaphorical sense. For example, the word 'president' means literally 'with a head'; this is a metaphorical extension of the 'head' that sits at the top of a human body.

Grammatical role is the function of a word in relation to other words in a sentence. Grammatical roles include subject, predicate, adjective, adverb, preposition, etc.

Colloquial Arabic, vernacular Arabic, dialectal Arabic, spoken Arabic are variants of what is termed as 'āmmiyyah or darija. They are the Arabic language variety used for day-to-day communication. Maghribi, Egyptian, Beduin, and Madani Arabic are examples of that variety (Elnagar *et al*. 2021a; Elnagar *et al*. 2021b).

3. Literature review

Most current language processing uses distributional representations of words and phrases as high dimensional vectors in a semantic space. That is why word embeddings are at the foundation of most applications in natural language processing. Yet, despite their popularity, their evaluation metrics have been criticized. Manzini *et al*. (2019) observed how the human tendency to make stereotypes gets amplified and biases get propagated when word embeddings are used. They applied Bolukbasi *et al*. (2016) method to remove bias components from texts and proposed a metric for the evaluation and quantification of bias in texts. Their method used Mean Average Cosine to measure the similarity between vectors.

Nissim *et al*. (2020) observed how vector spaces do encode human bias and they noted how some of the literature has found analogy metrics to be "deeply infused with human biases, like man is to computer programmer as woman is to homemaker", so they demonstrated that "the bias isn't necessarily (or at least not only) in the representations themselves, rather in the way we query them" (p. 5). In W2V, for example, the original two-pair analogy (A:B as C:D) (e.g., man is to king is as woman is to queen) necessitates that all four terms be distinct; thus, it prevents the return of man is to king as woman is to king (i.e., D is B).

Analogical reasoning is heavily utilized in the detection of language features, in morphological analysis, and in word sense disambiguation. Mikolov *et al*. (2013) developed the Google Analogy Test to evaluate the goodness of word embeddings by solving analogy questions with vector offsets; many researchers adopted this test whenever they wanted to assess embeddings and to uncover language patterns. One major limitation in this test was identified by Köper, Scheible, and im Walde (2015). They observed that such morphological richness as found in German, for instance, does indeed make the prediction of analogies more difficult. They demonstrated how the overall performance of continuous representations in German is lower than it is in English and that these representations "lack the ability to solve analogies of paradigmatic relations" (p. 44). Hence, synonymy relations are not easily identified.

Relations in analogy tests have also been questioned. Gladkova, Drozd, and Matsuoka (2016) are critical of the limited relations that they consist of. They assert, "to make any claims about a model being good at analogical reasoning, we need to show what types of analogies it can handle.

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

This can only be determined with a comprehensive test set" (p. 8–9). They are critical that most existing tests exhibit "unbalanced sets [of relations], and potentially high variation in performance for different relations (word-formation getting particularly little attention)" (p. 8).

We have also observed that non-English analogy tests are, for the most part, translations of English tests, a fact that compromises the discovery of patterns peculiar to each language. Zahran *et al.* (2015) translated Mikolov *et al.* (2013) Google Analogy Test into Arabic in verbatim. Elrazzaz *et al.* (2017) produced the first analogy test that consisted of authentic rather than translated Arabic words. They adapted the Google Analogy Test by removing six of the English relations because they were not relevant to Arabic, kept eight relations that they thought were relevant, and added only one new relation, the singular-dual, that they named 'pair'.

Köper *et al.* (2015) translated the Google Analogy Test into German but deleted the adjective-adverb relation because it did not exist in German. Ulčar *et al.* (2020) translated the English dataset into nine European languages but modified it such that they removed language and culture-specific categories. Their experimentation proved there to be "differences across languages and categories, and [there to be] a substantial room for improvement in creation of word embeddings that would better capture relations present in the language as distances in vector spaces" (p. 6). Khusainova, Khan, and Rivera (2019) developed a new dataset for the Tatar language that is based on Mikolov's English dataset. It consists of seven semantic and 27 syntactic categories that reflect the morphological richness of the Tatar language and culture (e.g., capital-republic in Russia; name-occupation; noun-derived adjective; five categories of grammatical case; verb mood, tense, and voice; etc.).

Gladkova *et al.* (2016) offered what they labeled 'a balanced test set' to systematically examine analogy-based detection of morphological and semantic relations in word embeddings. Their Bigger Analogy Test Set covers 40 linguistic relations in morphology and semantics. The morphology category consists of word pairs that exemplify both derivational and inflectional morphology. It consists of derivational prefixes (un-, over-, and re-); derivational suffixes (-less, -ly, -ness, -able, -er, -ation, and -ment); and inflectional forms (regular and irregular plurals, comparative and superlative adjectives, and participial and past verb forms). The semantic category consists of word pairs that exemplify lexical relations (i.e., hypernyms, hyponyms, meronyms, synonyms, and antonyms); and encyclopedic relations (i.e., geography, people, animals, and others). Each relation is represented by 50 unique word pairs: thus, yielding 2480 questions in each of the 40 categories, totaling 99,280.

The research reviewed thus far clearly highlights some inadequacies in the way word embeddings are evaluated. Mikolov *et al.* (2013) Google Analogy Test remains the most popular benchmark for the assessment of distributional representations. In the next section, we intend to make the case that Arabic word embeddings need to be more systematically assessed.

The use of word embedding models is quite popular in Arabic computational linguistic applications, AL-Smadi *et al.* (2017), Alkhatlan, Kalita, and Alhaddad (2018), Mohamed and Shokry (2022), Bounhas, Soudani, and Slimani (2020). Therefore, measuring the robustness of such embeddings is essential for producing effective, Arabic sentiment analysis, Altowayan and Elnagar (2017), Al-Ayyoub *et al.* (2019), Al-Smadi *et al.* (2019), Khalifa and Elnagar (2020), Nassif *et al.* (2021b), Farha and Magdy (2021), question and answering systems, Romeo *et al.* (2019), Einea and Elnagar (2019), clustering, AlMahmoud, Hammo, and Faris (2020), and text classification, Abbas *et al.* (2019), Orabi, El Rifai, and Elnagar (2020).

4. Proposed benchmark

Since Arabic is a morphologically introflexive, fusional language, and syntactically an inflectional language (Velupillai 2012), it is possible to produce close to half a million stems with a relatively small set of roots (5600–8000, depending on how and what you count). When inflected

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

for grammatical function, these stems produce millions of word forms. Good machine learning ought to be capable of utilizing the highly productive derivational and inflectional morphology to extract the language patterns hidden within. Hence, the evaluation gauge must be sophisticated enough to discriminate between an efficient and an inefficient word embedding.

Consequently, it is put forward here that the word analogy test be reflective of the nature of the language. It is proposed that the test capture the major features of noun and verb morphology, and its items be of a range of frequency, rather than of exclusively high frequency; for highly frequent patterns are necessarily easier to capture than patterns with low frequency. Therefore, our proposed analogy test has been constructed such that its categories cover (1) nouns and verbs, (2) the derivational, inflectional, and semantic dimensions, and (3) high-, medium-, and low-frequency items. It consists of three major analogy categories: morphosemantic, morphosyntactic, and semantic. We recognize that there are numerous dimensions left out for the sake of practicality, yet this test is capable of capturing a glimpse of how nouns and verbs are structured and conjugated and how they relate to one another semantically.

Below is an outline of the proposed analogy benchmark together with the justification for each category and subcategory in effort to maintain transparency and proper representation of the Arabic language.

Roots. The derivational nature of Arabic morphology centers around the root, and the morphological pattern, الوَزن الصَّرفِي al-wazn al-ṣarfī '*morphological measure*'. Some patterns are more productive than others. Alam (1983) counted the instances of types of roots in five classic dictionaries to find out the frequency of every type. He discovered that bilateral roots constitute only 1% of the roots in use, trilaterals 64%, quadrilaterals 33%, and quintlaterals 2%. The proposed word analogy test will include only trilateral roots since this type constitutes the majority of roots in the repertoire of the native speaker and because covering the two most frequent root types would double the size of the analogy test.

Roots, in terms of consonant type, are also two categories: all-consonantal الصَّحِيح al-ṣaḥīḥ and semivowel-inclusive-roots المُعتَلّ al-mu'tall. Yagi (2002) found all-consonantal roots to constitute three quarters of roots in the language.

To keep the size of the analogy test manageable, we decided to represent only major root categories and those that undergo morphophonemic transformation causing root radicals to get disguised. Therefore, we settled on 43 as the number of roots to include in our analogy test and decided the count of roots in each category in accordance with Yagi (2002) root types and frequencies. Our list correlates quite strongly with Yagi's root categories, $r(13) = 0.98$, $p = 000$.

4.1 Morphosemantic Indicators

Arabic derives its verbs and nouns by casting the root with its wholly abstract meaning into the mold of a morphological pattern with its own abstract meaning, such that the resulting stem would bring together the meanings of the root and pattern. The verb patterns are 27 but the most productive ones are 10, while the noun patterns are by Sibawayh and Ya'qub (1999) count 308 but the most productive are fewer than 60. Obviously, not all these morphological patterns can be covered in the analogy test, so we decided to have two verb patterns and three noun patterns to represent the noun and verb morphologies of Arabic. The decision of which verbal patterns to include was guided by two considerations:

(1) The morphological pattern has to rely on letter rather than diacritic manifestation; the purpose being that most Arabic texts are unvowelized, hence, it would not be feasible to

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

automatically distinguish between patterns. Take a morphological pattern like فَعَّلَ fa‘‘ala; it relies on the gemination diacritic (shadda) to be set apart from a pattern like فَعَلَ fa‘ala.

(2) The pattern has to be of significant frequency.

That is why we chose أَفعَلَ ’af‘ala and اِستَفعَلَ istaf‘ala; the earlier being of the highest frequency and the latter being the longest pattern and with one of the lowest frequencies. Furthermore, the verbal noun was selected for inclusion so that both verb and noun morphology would be represented. We also included the highly frequent active participle as a representative of derived nouns. Because nouns are often inflected for gender, we also thought it would be a good idea to represent that aspect of noun morphology even though gender is not strictly a morphosemantic category. Thus, the derivational dimension of Arabic morphology is reasonably well covered. As for the inflectional dimension, it is best captured by a sample of morphosyntactic paradigms.

4.2 Morphosyntactic Indicators

After roots get cast into morphosemantic patterns to derive nouns and verbs, they have to function in sentences. Their relation to one another and to other words in the sentence is often referred to as grammatical function. This grammatical function is marked by inflections, that is, prefixes and suffixes, that could alter the citation form of the word. Verbs, for instance, are inflected for tense, voice, and mood, and nouns are inflected for case, number, gender, and definiteness, but these inflections could cause morphophonemic alteration that camouflages the original word. Take the root *wfy* (وفي) *‘to keep a promise’* when cast into the pattern fa‘ala *‘did something’*, it produces the word wafā (وفى) *‘kept a promise’*. When conjugated with pronominal suffixes, the word would change form to wafat (وفت) *‘she kept promises’*; wafā (وفى) *‘he kept promises’*; wafū (وفوا) *‘they (masc.) kept promises’*; tafī (تفي) *‘she keeps promises’*; yafī (يفي) ‘he keeps promises’; and yafūna (يفون) *‘they (masc.) keep promises’*. A good word embedding ought to be able to capture the relationship between these different word forms. That is why the analogy test must include items that capture this inflectional nature of Arabic morphology. Because it is impractical to include all morphosyntactic alterations, our test includes only perfect and imperfect verbs conjugated for the third person feminine, hiya (هي) *‘she’*; the masculine singular, huwa (هو) *‘he’*; and the masculine plural, hum (هم) *‘they (masculine)’*. We ignored the dying dual and feminine plural because of their extremely low frequency.

Thus, words in our benchmark represent the Arabic word in its complexity: the abstract meaning in two root types, the all-consonantal and the semi-vowel inclusive; the lesser abstract morphosemantic template, represented by the morphological patterns أَفعَلَ ’af‘ala and اِستَفعَلَ istaf‘ala; and the grammatical inflections captured by the morphosyntactically marked word forms, the imperfect verb conjugated for the third person singular and plural feminine and masculine pronouns.

4.3 Semantic Indicators

To reflect the semantics of the Arabic language, the word analogy test must include lexical semantic categories: synonymy, antonymy, and hyponymy, and these must be derived from a corpus of authentic language use. That is why, it was decided that items in these categories will be drawn from Buckwalter and Parkinson (2011). This is a frequency dictionary of Arabic that rank-orders the 5000 words at the top of the frequency list of a corpus of 30 million words in size.

We selected items for the categories of our proposed benchmark such that they would be equally divided between nouns, verbs, and adjectives, and that they would be of diverse frequencies.

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

The lexical items included in the semantic indicators were selected from the Part of Speech (POS) Index of this dictionary. The selection was done in a disciplined manner; the first half of items in a semantic category would be drawn sequentially from the top of the frequency list, the second half would be selected at a 300-rank interval. Thus, half of the items in each of the semantic indicator lists were chosen in sequence from the top of the adjective, noun, and verb sections of the POS index. The other half was selected from the remaining ranks at an approximately 300-rank interval. So, when we compiled the list of adjectives, for example, we had to pick every adjective encountered at the top of the POS section of adjectives and the adjectives at the frequency ranks of 304; 598; 903; 1206; 1503; 1803; 2100; 2400; 2713; 3002; 3311; 3605; 3897; 4186; 4519; 4804; and 4993.

4.4 Composition of the Benchmark Dataset

In this section, we present our proposed Arabic benchmark dataset accepting that a word analogy test should be reflective of the nature of the language, such that test items be (1) inclusive of high, medium, and low frequency; (2) demonstrative of the morphosemantics of roots and patterns; (3) indicative of the morphosyntax of conjugated verbs and nouns; and (4) inclusive of the major semantic relations. Here is an outline of it.

The complete benchmark comprises 44 tests, each consisting of 40–50 word pairs. The full set of tests is available online on Github.[a] It is contained in three major categories:

(1) Morphosyntax indicators which are covered in the benchmark by: 13 files for the imperfect verb of the ṣaḥīḥ sālim (صحيح سالم) roots (CISS) conjugated from the past tense for the pronouns: Huwa, Huma, Hum, Hiya, HumaF, Hunna, Anta, Antuma, Antum, Anti, Antunna, Ana, Nahnu; 13 files for the imperfect verb of the waw-inclusive (CIWI) roots (صحيح معتل الواو) conjugated also from the past for the same set of pronouns; 2 files for the imperfect verb of the third person masculine and feminine plural (i.e., yataFa3aLuuna and yataFaa3aLna, respectively) conjugated from the singular form.

(2) Morphosemantics indicators which are expressed in 11 files: the patterns (أَفعَلَ) 'af'ala; (اِفتَعَلَ) ifta'ala; (اِنفَعَلَ) infa'ala; (اِستَفعَلَ) istaf'ala; (تَفَعَّلَ) tafa''ala; (تَفاعَلَ) tafā'ala; the Active-Participle; Verbal-Noun; Masculine-Feminine; Feminine-Singular-Plural; Masculine-Singular-Plural.

(3) Semantics indicators which are represented by five files: three of which for the lexical relations: Synonyms, Antonyms, Hyponyms, and one each for Capital-Cities, and Currencies.

The process of testing involves studying the word analogy relationship (i.e., $word_1$ to $word_2$ is like $word_3$ to x) between each pair of words, in the file, against the remaining pairs in the same file. The objective is to recover the word x as $word_4$. This process is repeated in each file for each benchmark indicator. This process is similar to what Mikolov *et al.* (2013) and Elrazzaz *et al.* (2017) did for evaluating word embeddings. In order to recover the missing word, it is suggested here that the top-5 matches be considered, rather than the top candidate as the case is for English. This is because Arabic derivational and inflectional nature allows the top candidate to appear in a word form that might differ from the targeted form by a one letter clitic or affix. Köper *et al.* (2015) lend support to our recommendation with their finding that the morphological richness of German makes the prediction of analogies more difficult; thus, the overall performance of embeddings in German is lower than their counterparts in English. Table 1 summarizes the composition of the

[a] https://github.com/elnagara/AREEB.

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

Table 1. Proposed benchmark metrics for Arabic embeddings

Inflectional morphology (MorphoSyntactic)	Derivational morphology (MorphoSemantics)	Semantics
Imperfect verb	Perfect verb	Noun/Verb/Adjective
ConjugatedImperfectSahihSalim-huwa	aF3aLa	Synonymy
ConjugatedImperfectSahihSalim-hiya	istaF3aLa	Hyponym
ConjugatedImperfectSahihSalim-hum	**Noun**	Antonymy
ConjugatedImperfectW-Inclusive-huwa	Verbal Noun	Capital City
ConjugatedImperfectW-Inclusive-hiya	Active Participial	Currency
ConjugatedImperfectW-Inclusive-hum	Masculine-Feminine	

proposed benchmark. It consists of components that capture Arabic verb morphology, noun morphology, and semantics. The verb morphology consists of two types: inflectional morphology that we refer to as 'morphosyntax' and derivational morphology that we refer to as 'morphosemantics'. The components that are designed to represent the morphosyntax of Arabic are targeted towards the imperfect as it contrasts with the perfect verb, and they show it in its multitude of forms as it is conjugated for the singular and plural, masculine and feminine, 1st person, 2nd person, and 3rd person pronouns. They also represent the two major types of trilateral roots: the all-consonantal ṣaḥīḥ sālim and the w-inclusive roots. Table 1 shows a sample of indicators that are discussed in detail here.

As for noun morphology, it is represented in the morphosemantics of perfect verbs and nouns. The perfect verb morphology is typified by the morphological patterns, aF3aLa and istaF3aLa. The noun morphology, on the other hand, is represented by the verbal noun, active participial, and by the masculine–feminine contrast.

Arabic semantics is depicted in the proposed benchmark by three basic sense relations: synonymy, hyponymy, and antonymy and by two of the components in Mikolov *et al.* (2013) and Elrazzaz *et al.* (2017): capital city and currency.

5. Experimental evaluation and discussion

To gauge the efficiency of the proposed benchmark, we will conduct analogy tests, and intrinsic and extrinsic evaluation. Before we commence, let us first introduce our benchmark datasets.

5.1 Datasets

We built 40 models of the datasets below to facilitate the evaluation of our benchmark. Evaluation is conducted here for all dichotomies in our benchmark using the embeddings produced by FT, W2V, and GloVe from the following four main datasets.

(1) **Historical Arabic Corpus (HAC)**
The books in the Historical Arabic Corpus (HAC) (Hammo *et al.* 2016) and the books of the Open Islamicate Texts Initiative (OpenITI) (Romanov and Seydi 2019); let this dataset be called BOOKS.
HAC is a historical Arabic corpus with around 45 million tokens, annotated with part-of-speech, root, and morphological pattern. Its content is categorized into 100-year time intervals and in terms of authorship, genre, and whether the text is a primary representation of the language of its epoch. The goal of the corpus is facilitation of the compilation of entries in a hypothetical dictionary on historical principles, Hammo *et al.* (2016).

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

The OpenITI is a multidisciplinary scholarly corpus of premodern Arabic texts that was compiled by researchers at academic institutions in the USA and Europe and funded by international philanthropies (Romanov, 2020). Its texts come from open-access online collections of premodern and modern Arabic texts, primarily from 'Shamela' and 'ShiaOnlineLibrary'. The total number of word types in it is over 745 million and 1346 million tokens.

(2) **News Articles (NEWS)**
Our large collection of news-wire articles from multiple sources, Einea, Elnagar, and Debsi (2019), and from the Masrawy dataset, Elnagar, Al-Debsi, and Einea (2020); let this collection be called NEWS.
NEWS collection is a large collection of Arabic news articles that can be used in different Arabic NLP tasks such as text classification and word embedding. The articles were collected using Python scripts written specifically for four popular news websites: AlKhaleej, AlArabiya, Akhbarona, and Masrawy. The first dataset (190K articles) has seven categories: Culture; Finance; Medical; Politics; Religion; Sports; and Technology, except for AlArabiya which does not have [Religion]. Masrawy is a huge dataset that contains 451K files that supports multi-class text classification. It has 24 categories.
For news-wire articles, we used the SANAD dataset, Einea *et al.* (2019), which is a large collection of Arabic news articles that were collected from three popular news websites: 'AlKhaleej', 'AlArabiya' and 'Akhbarona'. SANAD contains almost 200,000 articles. Similarly, the Masrawy dataset (`Masrawy.com`) consists of more than 450,000 news articles collected from the Masrawy portal.

(3) **Wikipedia (WIKI)**
Arabic Wikipedia documents; let this dataset be called WIKI. Wikipedia is an encyclopedia that compiles its content in a variety of languages. The Arabic content domain is a relatively rich one with more than one billion tokens, spanning several different areas of knowledge. We used it as a separate source for building wiki-based embedding models. We obtained the Arabic dump back in January 2020.

(4) **Reviews (REVIEWS)**
Our huge collection of Hotel Arabic Reviews dataset (HARD, Elnagar, Khalifa, and Einea 2018a) and our Book Reviews of Arabic Dataset (BRAD, Elnagar, Lulu, and Einea 2018b), both include dialectal Arabic and emojis. Let the combined datasets be called REVIEWS. HARD and BRAD are datasets whose texts were scrapped from `booking.com` and `goodreads.com` portals, respectively. The two datasets have more than 800K reviews in total.
The latest versions of HARD and BRAD datasets contain 500,000 and 700,000 reviews, respectively. While the HARD dataset represents hotel reviews that were collected from `Booking.com` website, BRAD reviews are about books and were collected from `GoodReads.com`. All reviews are expressed in Modern Standard Arabic as well as dialectal Arabic.

In addition, we used a dataset by Soliman, Eissa, and El-Beltagy (2017), which is a pre-trained word embedding available as open source; let it be called AraVec. We included the AraVec-Wikipedia model for comparison purposes. We used the above four datasets to construct experimental embedding models as explained in the following section.

As for our analogy tests, we used only a sample of test files to illustrate the capabilities of the benchmark. It is impractical to include all here; a sample of a paradigm is sufficient to demonstrate it. We selected the highest and the lowest frequency morphosemantic and morphosyntactic patterns. Researchers using our benchmark are not expected to use all testing files, a subset like what we demonstrate here is sufficient. Embeddings that work on the subset ought to work on

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

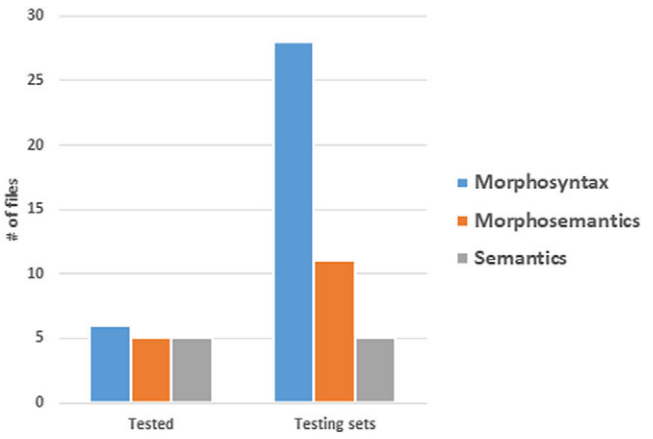

Figure 1. Complete benchmark indicators (Testing set) versus the subset used for experimentation (Tested set).

the whole benchmark as shown in the evaluation section. Figure 1 shows the distribution of test files per morphosyntactic, morphosemantic, and semantic indicator. The testing set reflects the total number of test files per benchmark indicator. The tested set shows the actual number of files used in the experiments conducted for evaluation of the benchmark. Specifically, the morphosyntax indicators have 28 files in the benchmark: 13 files for conjugated imperfect (CISS) ṣaḥīḥ sālim (صحيح سالم) roots, 13 files for conjugated imperfect waw-inclusive (CIWI) roots (صحيح معتل الواو). However, in this section we show the results of only 6 selected test files for the benefit of space-saving: CISS-hiya, CISS-hum, CISS-huwa, CIWI-hiya, CIWI-hum, and CIWI-huwa; this is sufficient to demonstrate the paradigm. The morphosemantic indicators have 11 files six of which are the morphosemantic patterns: (أَفعَلَ) ’af‘ala; (اِفتَعَلَ) ifta‘ala; (اِنفَعَلَ) infa‘ala; (اِستَفعَلَ) istaf‘ala; (تَفَعَّلَ) tafa‘‘ala; (تَفاعَلَ) tafā‘ala; Active-Participle; Verbal-Noun; Masculine-Feminine; Feminine-Singular-Plural; Masculine-Singular-Plural. For testing this indicator, we selected the most and least frequent morphosemantic patterns ((أَفعَلَ) ’af‘ala, (اِستَفعَلَ) istaf‘ala) and the rest of indicators. The third benchmark indicator is semantics, which has five files: Synonyms, Antonyms, Hyponyms, Capital-Cities, and Currency. We used all five files in testing this indicator.

5.2 Analogy Assessment

The above datasets will be used in the evaluation of word embedding models with the aid of analogy tests. Ever since Mikolov *et al.* (2013) such models have always been evaluated using analogy relations which rely heavily on the notion of similarity.

The goal of word similarity is to measure how well the notion of human perceived similarity is captured by word vector representation. Therefore, word similarity correlates the distance between word vectors and human perceived similarity. Word analogies are equivalent to word transformations that describe common semantic differences. In this work, we consider similarity to be part of analogy.

For all the experimental models discussed here, we included uni-gram as well as n-gram models (bi-gram and tri-gram). The embedding models are used to search for the nearest neighbors in the embedding space or as an embedding layer in a machine learning model in a supervised task (to be discussed in Subsection 5.4). Therefore, good embedding models must maintain relationships

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

Table 2. Top 3 results retrieved from the W2V-CBOW embedding of 'NEWS'

SN	Query	Top three results		
1	الأمم المتحدة	الامم المتحدة	المنظمة الدولية	منظمة الأمم المتحدة
2	السيدة عائشة	رضي الله عنها	أم سلمة	أم حبيبة
3	المسجد الأقصى	المسجد الأقصى المبارك	الحرم القدسي	الأقصى المبارك
4	فيس بوك	فيسبوك	موقع فيسبوك	واتس آب
5	سخيف	مضحك	ساذج	غير معقول
6	رائع	مميز	مبهر	مذهل
7	قبلة	وقبلة	وجهة رئيسية	مقصدا
8	مبروك	ومبروك	ألف مبروك	هارد لك
9	قمر	القمر	قمري	مسبار
10	كذا	كذا وكذا	فلان	الفلاني
11	هون	تشونج	سانغ	جونج

between words. To illustrate nearest neighbors and analogy relationships, we show a sample of query terms with their top 3 neighbors as retrieved from 'W2V-CBOW' models for the following sample datasets:

(1) 'NEWS' which is based on SANAD and Masrawy, Table 2.

(2) 'REVIEWS' which is based on HARD and BRAD, Table 4.

Tables 2 and 4 show the resulting top 3 neighbors for 11/12 query terms purposefully selected such that they would demonstrate unigram versus n-gram, dialectal versus standard, polysemous versus nonpolysemous dichotomies for the W2V-CBOW (W2V-CBOW) architecture when applied to the two datasets.

Table 2 shows the top three results from the embedding of the NEWS dataset, which constitutes nonacademic texts that the lay people interact with on a daily basis. Transliterations is provided in Table 3. Notice that the query is sometimes a uni-gram and sometimes a bi-gram. The returned results are similarly uni-gram or n-gram, regardless of the number of words in the query (e.g., 1, 7, and 8). The top three results are invariably relevant; either they are synonyms, near synonyms, or frequent collocates. If we consider the ambiguous query words, we would notice the following:

The query word in 7 could be read to mean 'kiss' or 'destination', but the NEWS embedding seems to have found the second reading dominant. In 8, 'mabrwk' is not interpreted as a proper noun; it is taken as 'congratulations', which is the prototypical meaning that first comes to mind in contemporary Arabic. In 9, the ambiguous query (i.e., moon/ a girl name) appears to have been interpreted literally as 'moon', that is why the results are 'the moon', 'lunar', and 'space probe'. In 10, the ambiguous query is taken as equivalent to the Latin adverb sic 'so, thus'. Interestingly, the spoken Gulf Arabic variant is in the top results. In 11, the query seems to have been read as an Asian named entity rather than as a noun meaning 'ease/peacefulness/shame'; or as a verb that means 'to make easy/to facilitate/to mitigate/ to disparage'.

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

Table 3. Transliterations of the top 3 results retrieved from the W2V-CBOW embedding of 'NEWS' (Table 2)

SN	Query	Top 3 Results		
1.	al-ʾumam al-muttaḥidah	al-ʾumam al-muttaḥidah	al-munaẓẓamah al-dawliyyah	munaẓẓamah al-ʾumam al-muttaḥidah
2.	al-sayyidah ʿāʾishah	raḍiya al-lhu ʿanhā	ʾummu salamah	ʾummu ḥabībah
3.	al-masjid al-ʾāqṣā	al-masjid al-ʾāqṣā al-mubārak	al-ḥaram al-qudsī	al-ʾāqṣā al-mubārak
4.	fīs būk	fīsbūk	mawqiʿ fīsbūk	wāts ʾāb
5.	sakhīf	muḍḥik	sādhaj	ghayr maʿqūl
6.	rāʾiʿ	mumayyaz	mubhir	mudhhil
7.	qiblah	waqublah	wujhah	maqṣadā
8.	mabrūk	wamabrūk	ʾalf mabrūk	hārd lak
9.	qamar	al-qamar	qamarī	misbār
10.	kadhā	kadhā wakadhā	fulān	al-fulānī
11.	hawn	tshwnj	sāngh	jūnj

Table 4. Top 3 results retrieved from the W2V-CBOW embedding of 'REVIEWS'

SN	Query	Top three results		
1	الأمم المتحدة	المنظمة الدولية	الحكومة اليمنية	المعارضة السورية
2	السيدة عائشة	رضي الله عنها	أم المؤمنين	السيدة خديجة
3	المسجد الأقصى	المسجد الاقصى	الأقصى	قبة الصخرة
4	فيس بوك	الفيس	الفيس بوك	الفيسبوك
5	سخيف	مستفز	ساذج	تافه
6	رائع	راائع	ممتاز	جميل
7	قبلة	بقبلة	بوسة	قبلة حنينة
8	مبروك	مبرووك	الف الف مبروك	مبروك عليك
9	قمر	نجم الدين	غدير	شروق
10	كذا	جذي	كدا	زي كذا
11	هون	هلأ	هنيك	هسا
12	☺☺	واو ☻☻	بابتسامة اوك	بصح ☺

Table 4 shows the results of querying the embedding of the REVIEWS dataset, the collection of hotel and book reviews. Transliterations is provided in Table 5. Since this is more reflective of everyday language as used in social media, many queries retrieved contemporary spoken language patterns such as vowel elongation in 6, 8, and 12; dialectal forms in 10, 11, and 12; and emojies in 12. The results of the homonymous query in 7 demonstrate the spoken nature of the dataset

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

Table 5. Transliterations of the top 3 results retrieved from W2V-CBOW embedding of 'REVIEWS' (Table 4)

SN	Query	Top 3 Results		
1.	al-ʾumam al-muttaḥidah	al-munaẓẓamah al-dawliyyah	al-ḥukūmah al-yamaniyyah	al-muʿāraḍah al-sūriyyah
2.	al-sayyidah ʿāʾishah	raḍiya al-lhu ʿanhā	ʾummu al-muʾminīn	al-sayyidah khadījah
3.	al-masjid al-ʾāqṣā	al-masjid al-ʾāqṣā	al-ʾāqṣā	qubbah al-ṣakhrah
4.	fīs būk	al-fīs	al-fīs būk	al-fīsbūk
5.	sakhīf	mustafiz	sādhaj	tāfih
6.	rāʾiʿ	rāāʾiʿ	mumtāz	jamīl
7.	qiblah	biqublah	bawsah	qublah ḥnayyinah
8.	mabrūk	mabruwwk	ʾalf ʾalf mbrwk	mabrūk ʿalayk
9.	qamar	najm al-dīn	ghadīr	shurūq
10.	kadhā	jidhī	kidā	zay kadhā
11.	hawn	hallaʾ	hunīk	hassā
12.	😊😊	wāāw 😊	biābtisāmah ʾūki	bṣaḥ 😊

really well. The query is no longer interpreted as 'a fulcrum or destination' but rather as the standard and colloquial words for 'kiss' and 'tender kiss'. The query in 9 ('moon') has no longer the strongest collocation with outer space (Table 2) but rather with metaphorical extension in the form of named entities (Najm Al-Dean, Ghadeer, and Shurouq). The query (هون, SN. 11) is no longer a noun or a verb relating to 'ease' (Table 2) or an Asian named entity but rather an adverb of time or place with different dialectal pronunciations. Indeed, the REVIEWS dataset is the most reflective of contemporary spoken Arabic.

Clearly, the embeddings are quite efficient since they, not only did not shy away from depicting n-grams for a uni-gram query, but also recognized the homonymy and polysemy of query terms as well as the dichotomous difference between standard and dialectal Arabic.

Now, let us turn to the evaluation of our proposed benchmark. Let us measure the efficiency of its performance intrinsically first, then extrinsically.

5.3 Intrinsic Assessment

We used the proposed benchmark to evaluate models of distributed word representations, measuring their ability to capture both the morphological and the semantic attributes of all datasets introduced in Subsection 5.1. Our indicators are expected to disclose to what degree these models are capable of extracting the morphological and semantic features of texts, the derivational and inflectional morphological characteristics, and noun and verb morphological attributes.

The results shown in this section are when the performance of models of distributed word representations of five datasets: HAC books; OpenITI, NewsArticles, Masrawy, and Ara-Vec. Three distributed word representation modeling algorithms are used: FT, GloVe, and W2V, all in Skip-Gram and Continuous Bag of Words architectures.

Figure 2 displays the morphosemantic evaluation of embeddings rendered for HAC using the SG and CBOW architectures of FT, GloVe, and W2V. Notice how the indicators can show the strength and weakness of the three embedding algorithms and their respective architectures. For ʾafala, the highly frequent morphosemantic pattern, FT-CBOW outperformed the other modeling algorithms by identifying around 50% of the query terms, two and a half times more than its next rival. FT's performance is most impressive when the least frequent morphosemantic pattern

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

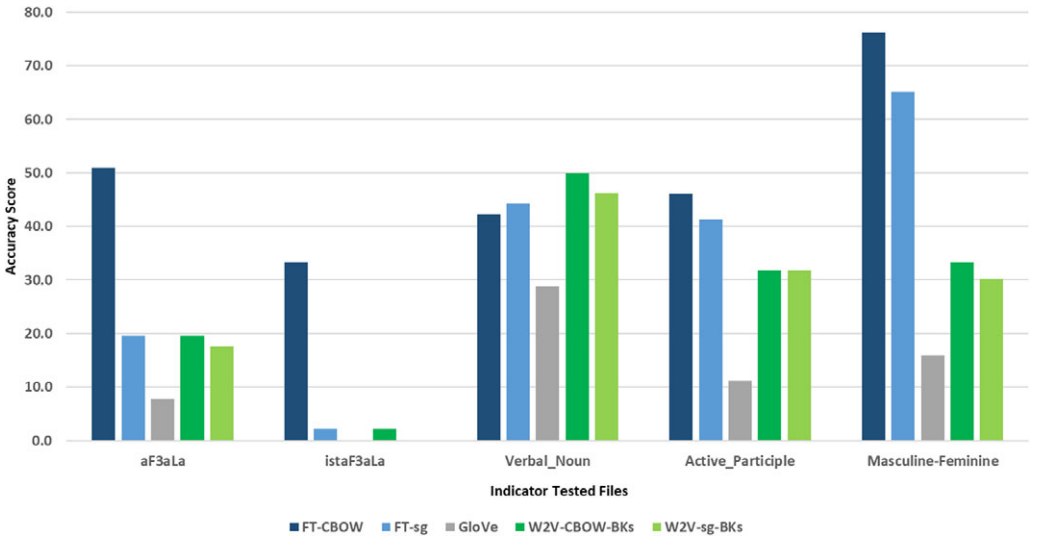

Figure 2. Morphosemantics in HAC embedding models.

is considered. It successfully spotted 33% of the instances of istafala in the top five query results; its next rival managed a success rate of less than 3%. In fact, the overall performance of FT surpassed all others, precisely because it is morphology-centric; it appears to capture best the nuances of Arabic morphosemantics.

This is not totally new but it partially corroborates the general consensus that FastText is best for morphological exploration as demonstrated in the verb derivational patterns of istafala, active participle, and in the gender contrast in 'masculine-feminine'. These results show Word2Vec as a worthy opponent in the noun derivation as manifested in verbal nouns. W2V outperforms FT in the detection of verbal nouns and rivals it in the detection of active participles. FT's continuous bag of words architecture is more efficient in derivational morphological exploration than its Skip-Gram architecture. Notice also that the frequency of occurrence of linguistic items has critical relevance for word representation. W2V Skip-Gram has barely been capable of detecting the relatively infrequent morphological pattern, istafala, while W2V Skip-Gram and GloVe have failed utterly in that. In highly frequent morphosemantic forms, the three embeddings, in their SG and CBOW iterations, have been quite good in detecting verbal nouns, active participles, and gender contrast.

Morphosemantics, however, might not be enough to draw this conclusion. Let us now check whether the embedding models would perform at the same level when morphosyntax and particularly inflectional morphology is targeted. We will use the proposed benchmark to discriminate between the three embeddings in the context of inflected word forms. Let us take the Conjugated Imperfect verb of the all-consonantal, ṣaḥīḥ sālim (صحيح سالم), (CISS), and the Conjugated Imperfect W-Inclusive, mutall, (CIWI), as cases in point.

Consider Figure 3 which shows the results of the three embedding models' performance on the morphosyntax in HAC. It should be noted that the axes in Figures 3–9 follow exactly what was presented in Figure 2.

All embeddings, whether in the SG or CBOW architectures, are capable of extracting morphosyntactic knowledge with reasonable efficiency. FT-CBOW is consistently the best in detecting conjugated word forms and GloVe tends to be the least efficient. CBOW is more capable than the SG architecture in revealing syntactically inflected word forms. Conjugated imperfect verbs that

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

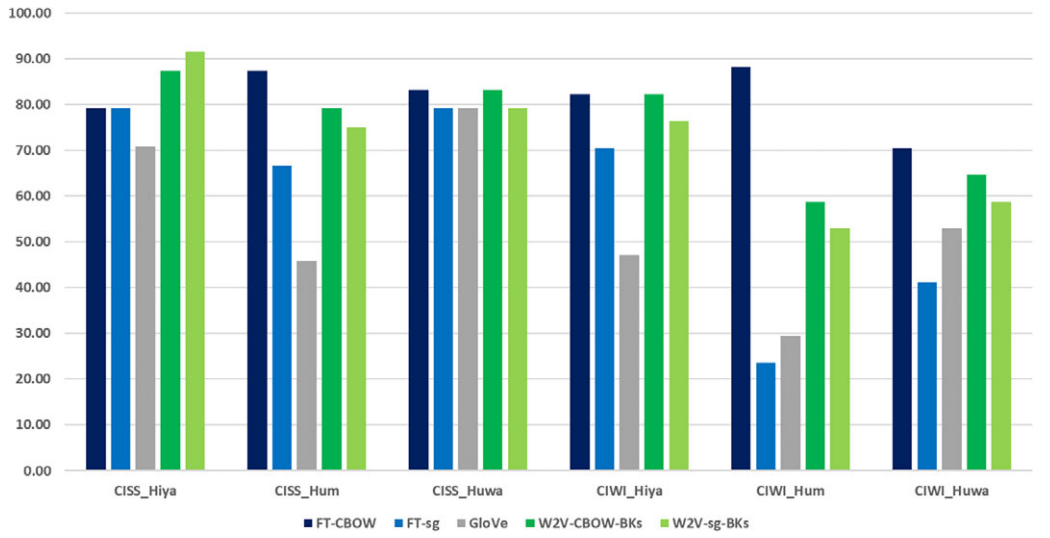

Figure 3. Morphosyntax in HAC embedding models.

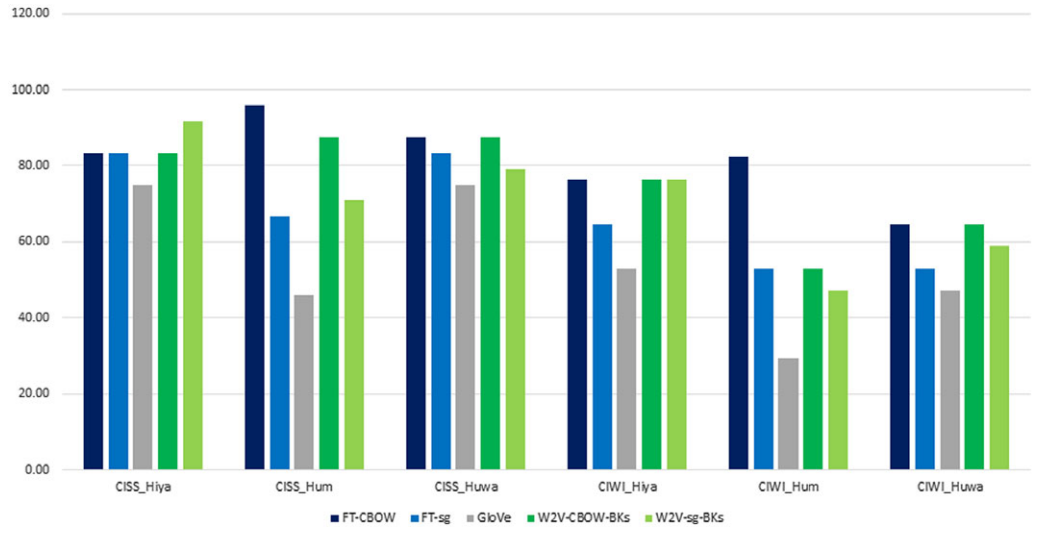

Figure 4. Morphosyntax in all-datasets embedding model.

are of the ṣaḥīḥ sālim (صحيح سالم) root type are better detected by the three main embeddings than conjugated imperfect verbs that are of the w-inclusive type.

Figure 4 is similar to Figure 3 but it depicts the morphosyntactic results for the embedding model of all-datasets combined rather than in just HAC. Although the patterns in the two figures are similar, the accuracy scores dropped a little in Figure 4. This is because we made the embedding models more general by combining them together despite the fact that the individual embeddings were created for specific NLP tasks. For example, developing a dialect classifier would not benefit as much from combining all-datasets as it would from the 'REVIEWS' embedding, which is predominantly dialectal.

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

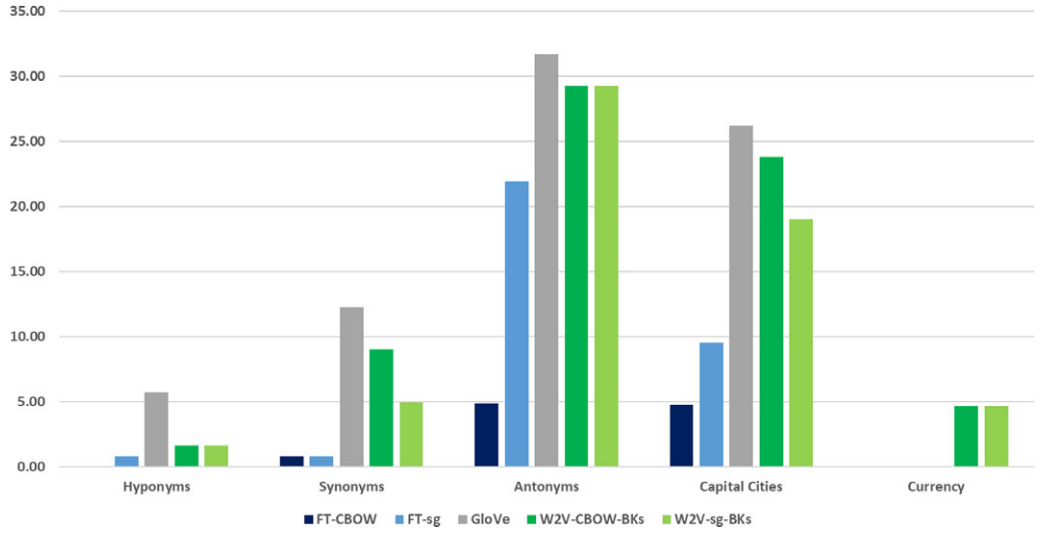

Figure 5. Semantics in HAC embedding models.

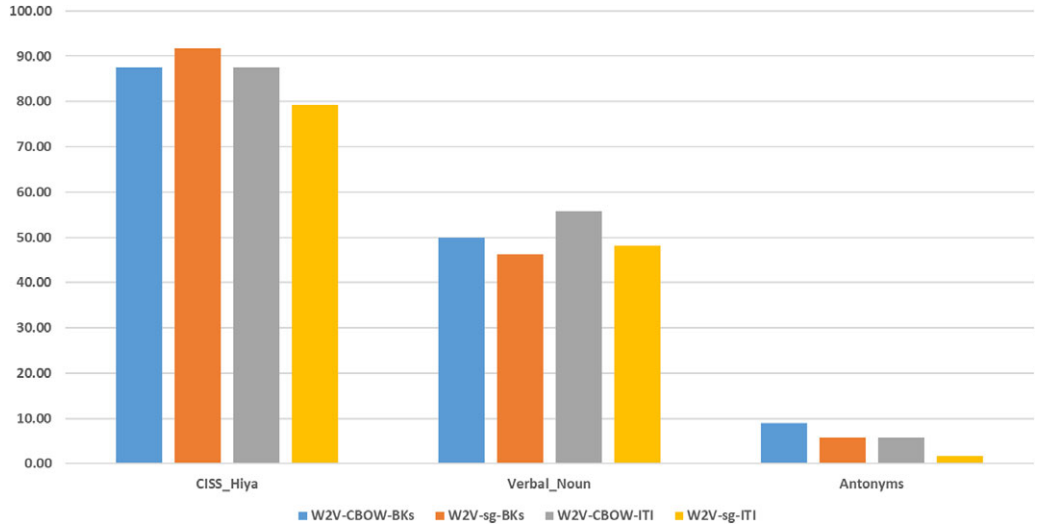

Figure 6. Word2Vec embeddings of HAC and OpenITI corpora.

Now, let us turn to the semantic dimension and see how well the three embedding models perform in representing texts in HAC. Figure 5 shows that GloVe is the best in detecting semantic features. This could be due to several reasons: (1) The overall performance of low-dimensional continuous word vector representations in such morphologically rich languages as German to be lower than in English (Köper *et al.* 2015); so they are expected to be of lower levels in Arabic as well. (2) They are generally less able to predict paradigmatic relations since "none of the vector spaces encodes deep semantic information reliably" Köper *et al.* (2015), p. 44. (3) The morphological richness of Arabic makes the prediction of analogies more difficult since the search space is larger with all the morphological variants and derivatives of a query term. Figure 5 also shows GloVe to be relatively better than the other two modeling algorithms in detecting semantic features.

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

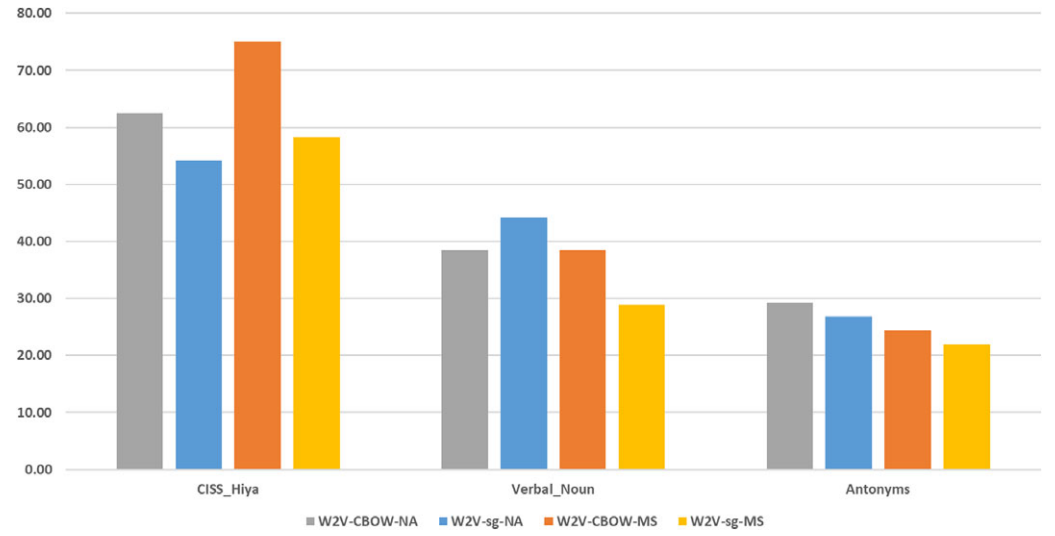

Figure 7. Word2Vec embeddings of Masrawy and other NewsArticles datasets.

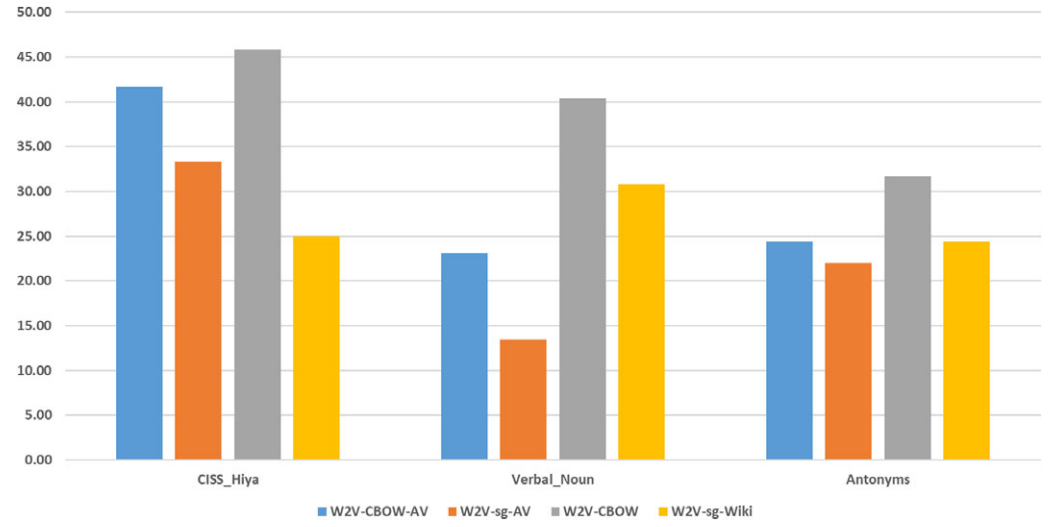

Figure 8. AraVec embedding versus our 'Wiki' embedding.

Let us now zoom in on one embedding model, say W2V, and examine to what extent would word embeddings be affected by corpus size. Subjecting the W2V embeddings of HAC and those in the similar in composition but larger OpenITI corpus, the proposed analogy benchmark revealed what is in Figure 6, that the magnitude of a corpus is not as critical as we might think, regardless of which architecture is adopted. Obviously, corpus size is important but beyond a certain point, it has diminishing returns in terms of quality of embeddings. Figure 6 shows how the quality of SG and of CBOW embeddings is only slightly affected by corpus size, regardless of whether we consider the morphosyntactics (exemplified by CISS_hiya, the all-consonantal imperfect verb conjugated for the third person singular feminine pronoun, 'hiya'), derivational morphology (represented by verbal nouns), or semantics (demonstrated by antonyms).

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

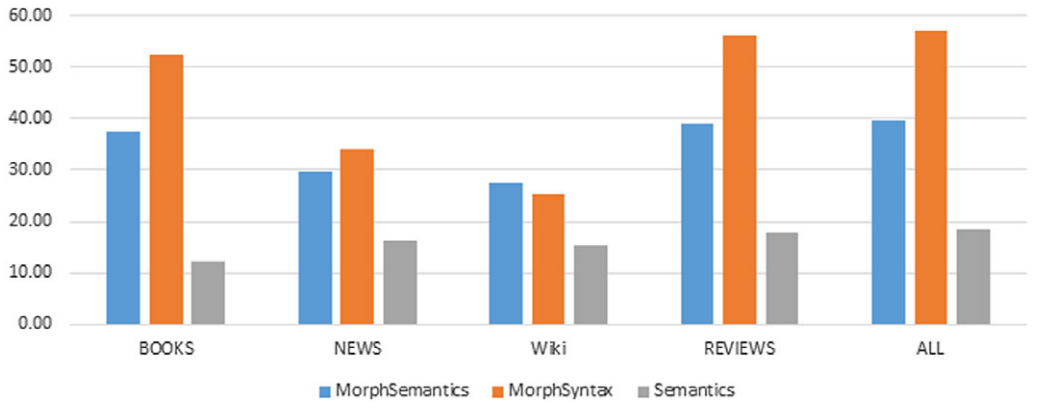

Figure 9. Overall performance on all test files grouped by benchmark categories.

HAC is only 45 million tokens compared to OpenITI's 1346 million tokens, yet its SG and CBOW embeddings reveal similar patterns.

The proposed benchmark shows the embeddings to be analogous. Figure 7 below displays the results of a comparison between, on the one hand, SG and CBOW embeddings in the huge single-source Egyptian news corpus, Masrawy, and on the other, those in a small, multiple-source news corpus that we compiled from a variety of news portals. Notice that the morphosyntactic component of our benchmark, illustrated by 'CISS_Hiya', reveals CBOW to be superior when the corpus is single-source and big than when it is multi-source and small. This conclusion is soon invalidated when meaning is taken into account. Consider the derivational morphosemantic Verbal Noun indicator and the semantic component of Antonyms. SG performs best when the corpus is small and when derivational morphology is in focus, as exemplified by verbal nouns. This, however, is invalidated when the morphosyntactic and semantic components of the benchmark are taken into account. It also appears that the use of a small corpus of multiple sources as in our NewsArticles collection vis-a-vis a huge single source corpus (i.e., Masrawy) has better coverage for morphosemantics as exemplified by verbal nouns and for semantics as represented by antonyms. This implies that the heterogeneity of a corpus enhances its semantic representation and makes up for the smallness of its corpus size.

Next, we discuss the effect of normalization on embeddings. The proposed benchmark can shed light on this issue, as demonstrated in Figure 8. With it, we evaluated respectively the SG and CBOW Word2Vec embeddings in two versions of the Arabic Wikipedia, a version with the text normalized and the other without normalization. Normalization, a widely adopted practice in Arabic computational linguistics, is the process of unifying the orthography of some Arabic characters. Namely, alif forms [آ ، إ ، أ ، ا] to [ا], hamza forms [ء ، ؤ ، ئ] to [ء], hā and tā marbūṭah [ة ، ه] to [ه], and yā and 'alif maqṣūrah [ي ، ى] to [ى]. Aravec did implement normalization but we did not because we thought normalization could affect the contextual meaning of such words as فَأر fa'r *'mouse'* and فار fār *'boiled'* or كرة kurah *'ball'* and كره kurh *'hatred'*.

The results show that our non-normalized texts (depicted in the two bars on the right in each cluster in Figure 8) give slightly better results than the normalized texts of Ara-Vec and that the CBOW embeddings are of better quality than those of SG irrespective of text normalization. Clearly then, word representation is not seriously hampered by the lack of text normalization.

Having demonstrated above the discriminatory power of the proposed benchmark on the sample test-files, consider Figure 9 to see the overall performance of the full benchmark (test files) grouped by morphosyntactic, morphosemantic, and semantic categories. In general, the embeddings performed better on the morphosyntactic indicators than they did on the meaning-based

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

morphosemantic and semantic indicators. This is to be expected as syntax has orthographic manifestation while the semantic ones do not.

With intrinsic evaluation, the overall test results have demonstrated that FT-based embedding produced the most favorable results. How would it perform extrinsically?

5.4 Extrinsic Assessment

Following Elrazzaz *et al*. (2017), we perform extrinsic evaluation of all proposed embedding models on two main tasks from NLP: Arabic Named Entity Recognition (ANER) and Text Classification (TC).

For ANER, the objective is to test all embedding models on their ability to detect Persons (PER), Organizations (ORG), and Locations (LOC). The aim of the TC task is to classify a given text in terms of five categories: Finance, Medicine, Politics, Sports, and Technology.

To carry out the two tasks, we implemented three deep learning networks using two machine learning approaches: Convolution Neural Networks (CNN) and Recurrent Neural Networks (RNN). For RNN, we experimented with both the Bidirectional Long Short-Term Memory (BiLSTM) layer and Bidirectional Gated Recurrent Unit (BiGRU). We selected these three models because of their good performance on similar tasks (Al Qadi *et al*. 2019; Al Qadi *et al*. 2020; Elnagar *et al*. 2020; Nassif, Darya, and Elnagar 2021a; El Rifai, Al Qadi, and Elnagar 2022).

The function of CNNs is learning spatial features of the data, and then convoluting down to a smaller subset of the data while trying to learn more features from the already learned data. CNNs utilize a special layer called the pooling layer, which combines multiple related inputs into one based on some specific rules. A max-pooling layer of size 2×2 would get the maximum value of a 2×2 window and discard the remaining three inputs. An average-pooling works the same way but uses the average instead of the max.

RNNs are designed to work best with sequential data, or data that changes over time, such as textual or speech data. Unlike other neural networks, RNNs can process information in a bidirectional fashion in order to allow for learning information from the previous as well as the next states.

The difference between LSTM networks and RNNs is the ability of an LSTM network to remember information from layers that are too far behind, such as the case of sentences in a paragraph. LSTM networks have a forget gate, as well as an update gate. As the name suggests, the forget and update gates determine whether to pass the current information forward or to discard them. On the other hand, GRUs vary from LSTM units by utilizing update gates and reset gates. The gates' tasks are to determine the amount of information from previous layers to be either moved onto the next layers or discarded. Both RNN models listed above can also be wrapped around with a Bidirectional wrapper, giving us 2 new models. Namely, BiGRU and BiLSTM. Both models are composed of 1 BiRNN layer. The reason for implementing Bidirectionality is because of the nature of text, where each word is defined by the preceding and the proceeding words. Bidirectional wrappers allow the layers to go over the data in both directions, resulting in a vector that is 2 times as big as a uni-directional layer.

In the ANER task, the goal is to label each word in each sequence using one of the following 3 labels: PER, LOC, and ORG, which represent different Named Entity classes. This task was trained and tested on a dataset of size 2400 sentence from the "ANERcorp" dataset, which is a manually annotated corpus in Arabic for ANER tasks, Benajiba, Rosso, and Benedruiz (2007).

The embedding layer of the three proposed deep learning models was used to test all constructed embedding models. The performance of all embedding models on the ANER task is shown in Figures 10 and 11.

Similarly, we tested the performance of the embedding model on the TC task in which the goal is to label each document by a single label from the five labels. This task was trained and tested on a dataset of size 17,500 document from the "arabiya" dataset, which is a manually annotated

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

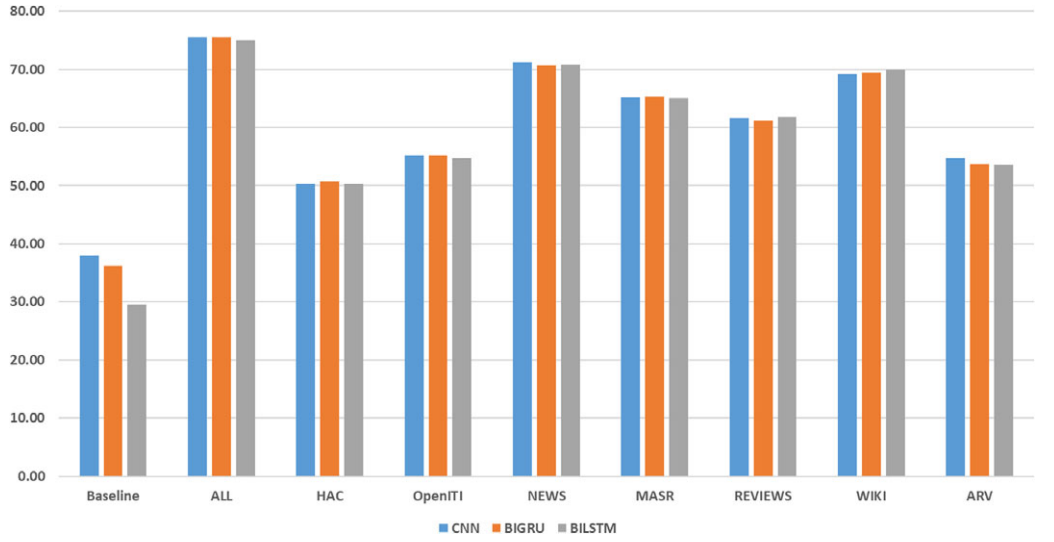

Figure 10. ANER accuracy results per dataset for all embeddings.

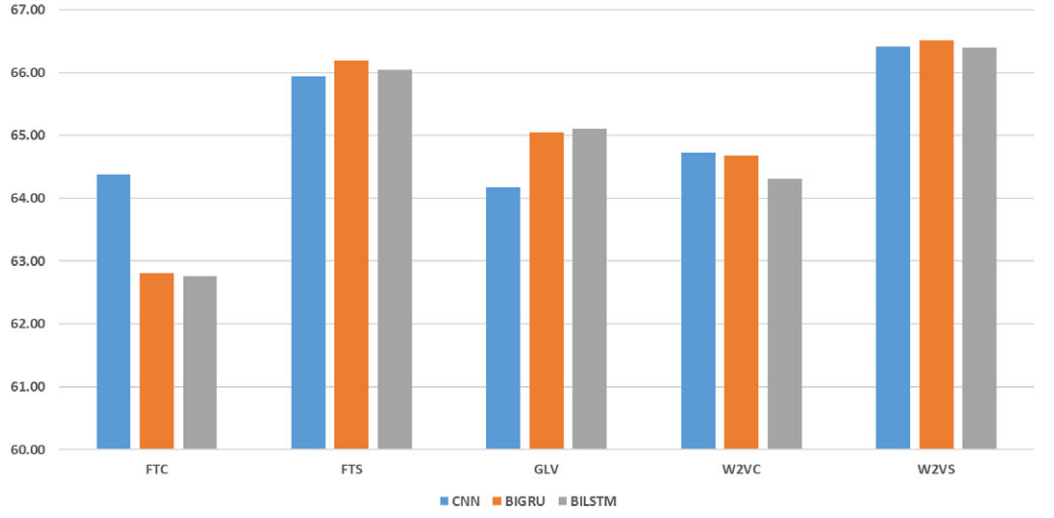

Figure 11. ANER accuracy results per embedding for all datasets.

corpus from 'Arabiya' news portal, Einea *et al.* (2019). The performance of all embedding models on the text classification task is shown in Figures 12 and 13.

With extrinsic evaluation, the overall test results indicate that W2V skip-gram embedding produced the most favorable results. However, FastText skip-gram nudges W2V so closely (see Figures 11 and 13). The skip-gram architecture outperforms the CBOW architecture for Arabic. As for the best dataset-based embeddings to use, it is all dependent on the desired NLP task. For example, for ANER, the all-datasets embedding is the best (Figure 10) since this embedding provides better coverage of names of organizations, people, locations, etc. However, for the text classification task, NEWS and WIKI-based embeddings provide the best results (Figure 12) since the texts they model are homogeneous (all written in MSA). In short, an NLP task would dictate which dataset-embedding to use.

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

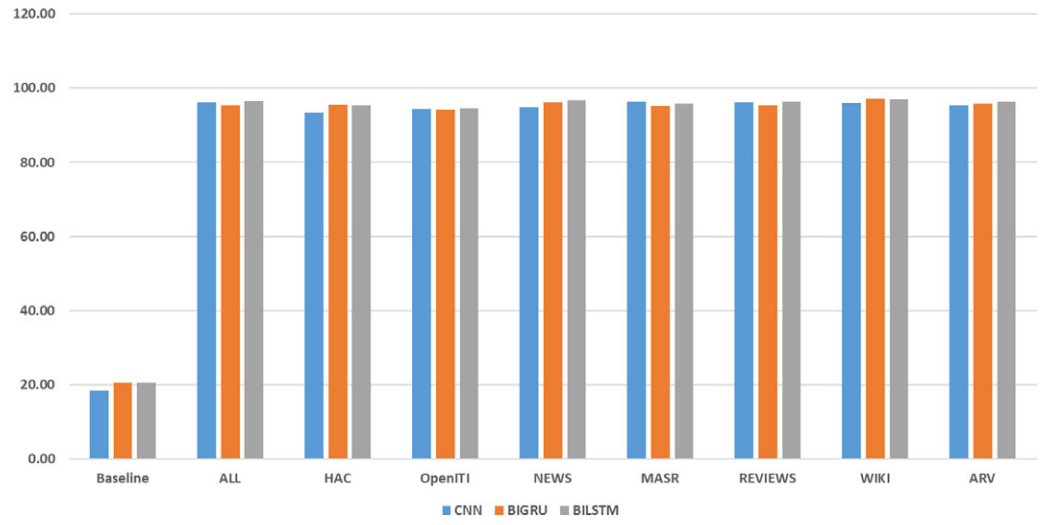

Figure 12. TC accuracy results per dataset for all embeddings.

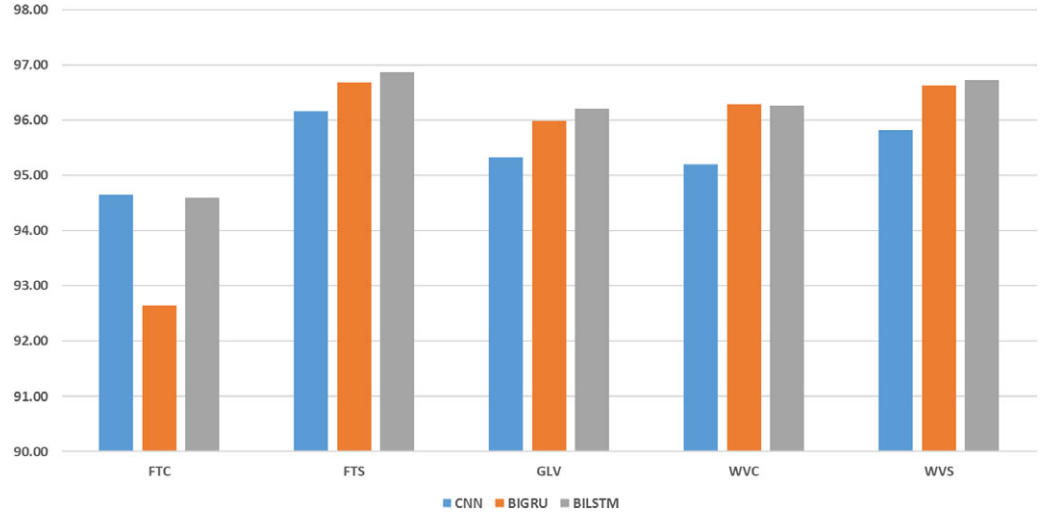

Figure 13. TC accuracy results per embedding for all datasets.

For text classification, we detail the CNN deep learning model, which used 356,869 trainable parameters. This network consists of several hidden layers including embedding, three layers of conv1d, max pooling, dropout, and dense. The number of epochs is set to 25 with an early stopping option. The experiments needed 12 epochs to report best accuracy scores. The `batch_size`=128. The `upper_limit` is 5000 characters, and `layer_outputs` is set to 128. We used a similar network for the ANER task.

6. Conclusion

We have presented here a benchmark for the evaluation of Arabic word embeddings after showing the inadequacy of current benchmarks, as they fail to take into account the root-based, introflexive, fusional morphology, and inflectional syntax of the language.

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

These features give it facility for the generation from several thousand roots, of millions of word forms and hundreds of thousands of word types. Our indicators evaluate distributional semantic models of this language by being attuned to its semantic, morphological, and syntactic intricacies.

The benchmark we proposed only partially echo Mikolov's benchmark; they are not a slavish translation of it. They have been designed such that they could be used to evaluate models of both Contemporary Arabic and Classical Arabic; they cover relations that are common to both varieties of this language.

Selection criteria of analogy items have been made transparent and truly reflective of the derivational and inflectional nature of the language. They capture the major features of nouns and verbs; derivational and inflectional morphology; high-, medium-, and low-frequency patterns and lexical items; and morphosemantic, morphosyntactic, and semantic dimensions of the language.

These indicators have been put to test both intrinsically and extrinsically by using them in the assessment of embeddings that were produced by FT, GloVe, and W2V in both SG and CBOW iterations. Their modeling of the distributed word representations of HAC books; OpenITI, NewsArticles, Masrawy, and Ara-Vec has revealed that FastText is the most suitable for Arabic morphosemantic and morphosyntactic investigations and that Word2Vec and GloVe are most suitable for semantic inquiry.

It has also demonstrated that corpus size has a point of diminishing returns, that heterogeneity of corpus content could compensate for the smallness of corpus size, and that graphemic normalization of Arabic texts creates polysemy and homonymy without compensatory gains in results.

The reliability and effectiveness of all the embedding models, that we have constructed, were demonstrated by analogy tests as well as two popular NLP tasks: named-entity recognition and text classification.

References

Abbas M., **Lichouri M.** and **Zeggada**, **A.** (2019). Classification of arabic poems: From the 5th to the 15th century. In Cristani, M., Prati, A., Lanz, O., Messelodi, S. and Sebe, N. (eds), *New Trends in Image Analysis and Processing – ICIAP 2019*. Springer International Publishing, pp. 179–186.

Al-Ayyoub M., **Khamaiseh A.A.**, **Jararweh Y.** and **Al-Kabi M.N.** (2019). A comprehensive survey of arabic sentiment analysis. *Information Processing & Management* 56(2), 320–342. Advance Arabic Natural Language Processing (ANLP) and its Applications.

Al Qadi L., **El Rifai H.**, **Obaid S.** and **Elnagar A.** (2019). Arabic text classification of news articles using classical supervised classifiers. In *2019 2nd International Conference on new Trends in Computing Sciences (ICTCS)*. IEEE, pp. 1–6.

Al Qadi L., **El Rifai H.**, **Obaid S.** and **Elnagar A.** (2020). A scalable shallow learning approach for tagging arabic news articles. *Jordanian Journal of Computer and Information Technology (JJCIT)* **6**(3), 263–280.

Al-Smadi M., **Al-Ayyoub M.**, **Jararweh Y.** and **Qawasmeh O.** (2019). Enhancing aspect-based sentiment analysis of arabic hotels' reviews using morphological, syntactic and semantic features. *Information Processing & Management* 56(2), 308–319. Advance Arabic Natural Language Processing (ANLP) and its Applications.

AL-Smadi M., **Jaradat Z.**, **AL-Ayyoub M.** and **Jararweh Y.** (2017). Paraphrase identification and semantic text similarity analysis in arabic news tweets using lexical, syntactic, and semantic features. *Information Processing & Management* 53(3), 640–652.

Alam Y.M. (1983). *al-Mujam al-Arabi: Dirasa Ihsaiya li-Dawaran al-Huruf fi al-Judhur al-Arabiya*. Thesis, Damascus University.

Alkhatlan A., **Kalita J.** and **Alhaddad A.** (2018). Word sense disambiguation for arabic exploiting arabic wordnet and word embedding. *Procedia Computer Science* **142**, 50–60. Arabic Computational Linguistics.

AlMahmoud R.H., **Hammo B.** and **Faris H.** (2020). A modified bond energy algorithm with fuzzy merging and its application to arabic text document clustering. *Expert Systems with Applications* **159**, 113598.

Altowayan A.A. and **Elnagar A.** (2017). Improving arabic sentiment analysis with sentiment-specific embeddings. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 4314–4320.

Bakarov A. (2018). A survey of word embeddings evaluation methods. CoRR, abs/1801.09536.

Benajiba Y., **Rosso P.** and **Benedruiz J.M.** (2007). Anersys: An arabic named entity recognition system based on maximum entropy. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pp. 143–153.

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

Bolukbasi T., **Chang K.-W.**, **Zou J. Y.**, **Saligrama V.** and **Kalai A.T.** (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in Neural Information Processing Systems, pp. 4349–4357.

Bounhas I., **Soudani N.** and **Slimani Y.** (2020). Building a morpho-semantic knowledge graph for arabic information retrieval. *Information Processing & Management* **57**(6), 102124.

Buckwalter T. and **Parkinson D.L.** (2011). *A Frequency Dictionary of Arabic: Core Vocabulary for Learners*. Routledge Frequency Dictionaries. London, New York: Routledge.

Einea O. and **Elnagar A.** (2019). Predicting semantic textual similarity of arabic question pairs using deep learning. In *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*. IEEE, pp. 1–5.

Einea O., **Elnagar A.** and **Debsi R.A.** (2019). Sanad: Single-label arabic news articles dataset for automatic text categorization. *Data in Brief* **25**, 104076.

El Rifai H., **Al Qadi L.** and **Elnagar A.** (2022). Arabic text classification: The need for multi-labeling systems. *Neural Computing and Applications* **34**(2), 1135–1159.

Elnagar A., **Al-Debsi R.** and **Einea O.** (2020). Arabic text classification using deep learning models. *Information Processing & Management* **57**(1), 102–121.

Elnagar A., **Khalifa Y.S.** and **Einea A.** (2018a). *Hotel Arabic-Reviews Dataset Construction for Sentiment Analysis Applications*. Computational Intelligence book series (SCI, Vol. 740), Springer International Publishing, pp. 35–52.

Elnagar A., **Lulu L.** and **Einea O.** (2018b). An annotated huge dataset for standard and colloquial arabic reviews for subjective sentiment analysis. *Procedia Computer Science* **142**, 182–189. Arabic Computational Linguistics.

Elnagar A., **Yagi S.**, **Nassif A.B.**, **Shahin I.** and **Salloum S.A.** (2021a). Sentiment analysis in dialectal arabic: A systematic review. In *International Conference on Advanced Machine Learning Technologies and Applications*. Springer, pp. 407–417.

Elnagar A., **Yagi S.M.**, **Nassif A.B.**, **Shahin I.** and **Salloum S.A.** (2021b). Systematic literature review of dialectal arabic: Identification and detection. *IEEE Access* **9**, 31010–31042.

Elrazzaz M., **Elbassuoni S.**, **Shaban K.** and **Helwe C.** (2017). Methodical evaluation of arabic word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 454–458.

Farha I.A. and **Magdy W.** (2021). A comparative study of effective approaches for arabic sentiment analysis. *Information Processing & Management* **58**(2), 102438.

Gladkova A., **Drozd A.** and **Matsuoka S.** (2016). Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pp. 8–15.

Hammo B., **Yagi S.**, **Ismail O.** and **Abushariah M.A.M.** (2016). Exploring and exploiting a historical corpus for arabic. *Language Resources and Evaluation* **50**(4), 839–861.

Khalifa Y. and **Elnagar A.** (2020). Colloquial arabic tweets: Collection, automatic annotation, and classification. In *2020 International Conference on Asian Language Processing (IALP)*. IEEE, pp. 163–168.

Khusainova A., **Khan A.** and **Rivera A.R.** (2019). Sart-similarity, analogies, and relatedness for tatar language: New benchmark datasets for word embeddings evaluation. arXiv preprint arXiv:1904.00365.

Köper M., **Scheible C.** and **im Walde S.S.** (2015). Multilingual reliability and "semantic" structure of continuous word spaces. In *Proceedings of the 11th International Conference on Computational Semantics*, pp. 40–45.

Manzini T., **Yao Chong L.**, **Black A.W.** and **Tsvetkov Y.** (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics, pp. 615–621.

Mikolov T., **Chen K.**, **Corrado G.** and **Dean J.** (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013, Workshop Track Proceedings*.

Mohamed E.H. and **Shokry E.M.** (2022). Qsst: A quranic semantic search tool based on word embedding. *Journal of King Saud University - Computer and Information Sciences* **34**(3), 934–945.

Nassif A.B., **Darya A.M.** and **Elnagar A.** (2021a). Empirical evaluation of shallow and deep learning classifiers for arabic sentiment analysis. *Transactions on Asian and Low-Resource Language Information Processing* **21**(1), 1–25.

Nassif A.B., **Elnagar A.**, **Shahin I.** and **Henno S.** (2021b). Deep learning for arabic subjective sentiment analysis: Challenges and research opportunities. *Applied Soft Computing* **98**, 106836.

Nissim M., **van Noord R.** and **van der Goot R.** (2020). Fair is better than sensational: Man is to doctor as woman is to doctor.

Orabi M., **El Rifai H.** and **Elnagar A.** (2020). Classical arabic poetry: Classification based on era. In *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*. IEEE, pp. 1–6.

Romanov M. and **Seydi M.** (2019). OpenITI: A Machine-Readable Corpus of Islamicate Texts.

Romeo S., **Da San Martino G.**, **Belinkov Y.**, **Barrón-Cedeño A.**, **Eldesouki M.**, **Darwish K.**, **Mubarak H.**, **Glass J.** and **Moschitti A.** (2019). Language processing and learning models for community question answering in arabic. *Information Processing & Management* **56**(2), 274–290. Advance Arabic Natural Language Processing (ANLP) and its Applications.

Schluter N. (2018). The word analogy testing caveat. In Walker M.A., Ji H. and Stent A. (eds), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*

https://doi.org/10.1017/S1351324922000444 Published online by Cambridge University Press

NAACL-HLT, New Orleans, Louisiana, USA, June 1–6, 2018, Volume 2 (Short Papers). Association for Computational Linguistics, pp. 242–246.

Sibawayh A.i.U. and **Ya'qub I.** (1999). *al-Kitab*. Dar al-Kutub al-Ilmiyah.

Soliman A.B., **Eissa K.** and **El-Beltagy S.R.** (2017). Aravec: A set of arabic word embedding models for use in arabic nlp. *Procedia Computer Science* **117**, 256–265. Arabic Computational Linguistics.

Ulčar M., **Vaik K.**, **Lindström J.**, **Dailidėnaitė M.** and **Robnik-Šikonja M.** (2020). Multilingual culture-independent word analogy datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association, pp. 4074–4080.

Velupillai V. (2012). *An Introduction to Linguistic Typology*. Amsterdam: John Benjamins Publishing Company.

Yagi S.M. (2002). Computerizing arabic morphology. *International Journal of Arabic-English Studies* **3**(1), 153–168.

Zahran M.A., **Magooda A.**, **Mahgoub A.Y.**, **Raafat H.**, **Rashwan M.** and **Atyia A.** (2015). Word representations in vector space and their applications for arabic. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pp. 430–443.

Cite this article: Yagi S, Elnagar A and Fareh S (2023). A benchmark for evaluating Arabic word embedding models. *Natural Language Engineering* **29**, 978–1003. https://doi.org/10.1017/S1351324922000444