Research Article **Open Access**

Stephen Taylor* and Tomáš Brychcín

The representation of some phrases in Arabic word semantic vector spaces

https://doi.org/10.1515/comp-2018-0017
Received July 20, 2018; accepted November 12, 2018

Abstract: We demonstrate several ways to use morphological word analogies to examine the representation of complex words in semantic vector spaces. We present a set of morphological relations, each of which can be used to generate many word analogies.

1. We show that the difference-vectors for pairs which have the same relation to each other are similarly aligned.
2. We suggest that addition of difference-vectors is a useful phrase-building operator.
3. We propose that pairs in the same relation may have similar relative frequencies.
4. We suggest that homographs, which necessarily have the same semantic vectors, can sometimes be separated into different vectors for different senses, using frequency estimates and alignment constraints obtained from word analogies.
5. We observe that some of our analogies seem to be parallel, and might be combined.

We use Arabic words as a case study, because Arabic orthography includes verb conjugations, object pronouns, definitive articles, possessive pronouns, and some prepositions in single word-forms. Therefore, a number of short phrases, built up of easily perceived constituents, are already present in stock semantic spaces for Arabic available on the web.

Similar phrases in English would require including bigrams or trigrams as lemmas in the word embedding, although English derivational morphology allows for other relationships in standard semantic spaces which Arabic does not, for example negation.

We make our corpus of morphological relations available to other researchers.

Keywords: phrase semantic vectors, word analogies, word embeddings, Arabic

1 Introduction

Semantic spaces, in which the usage in context, and by extension, the meaning of a word, is represented by a vector of real numbers [1], have become popular sources for word features in a variety of different natural processing tasks [2–4].

The idea that the meaning of a word is determined by its usage goes back at least to 1954 [5]. Deerwester [6] applied this idea to summarizing the meaning of a document for information retrieval, by first building for each document a $|\boldsymbol{V}|$-dimensional vector of frequency-statistics of words from the vocabulary $\boldsymbol{V}$, and then using Singular Value Decomposition to project these vectors into a smaller-dimensional space. The technique is called Latent Semantic Analysis, (LSA) and can be applied to create vectors for large documents, paragraphs, sentences, or single words. The resulting vectors can be compared with a distance metric to see whether two sentences or two words are similar in meaning, and how similar they are.

Today's semantic spaces determine word usage relative to other words in the nearby context, which is slightly different than determining the documents in which they are used; and general corpora are used, as the objective is to create word semantic features, and not to locate related documents.

Since 2003 [7], the algorithms for developing word vectors have been based on learning to predict the context, rather than counting it. Several fast algorithms have been developed for building word semantic spaces from large corpora including Mikolov's `word2vec` methods [8] `SkipGram` and `CBOW`, `GloVe` [9], and Bojanowski's [10] `fastText` code, and semantic spaces generated with them

***Corresponding Author: Stephen Taylor:** Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic;
E-mail: stepheneugenetaylor@gmail.com
Tomáš Brychcın: Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic; E-mail: brychcin@kiv.zcu.cz

are available on the web for many languages, including Arabic [8, 11, 12].

In [8], Mikolov also introduced the idea of using word analogies as a test of the quality of a semantic space, and made available his set of word analogies for English. In [13], additional analogies were added. Levy and Goldberg [14] examined the vector arithmetic used in Mikolov's examples, and showed that it is also applicable to LSA vectors. Since the non-reduced versions of these vectors can be examined to see exactly which context words participate in cosine distance calculations, it is possible with some hand-waving to explain how the various components of meaning are modified by vector arithmetic.

In recent years, some authors have focussed on extending word semantic spaces to create vector representation of phrase meanings. Vector spaces distinct from word spaces include `Paragraph2vec` [15], `Skip-thought` [16], `FastSent` [17], `Charagram` [18], all of which were devised to provide application-independent fixed-length encodings of arbitrary-length word sequences; and several encodings based on intermediate stages of neural networks trained for specific tasks, e.g. machine translation [19]. Conneau et al. [20] compare 24 different variants of vector encoders for sentences.

More relevantly to our phrases, which share the semantic space with words, are collocations, like *United States* الولايات المتحدة. These are two separate words, whose combined meaning is different than the combination of the meanings of the two words. However, in this case, the combined meaning seems to be similar to meanings of words like *France*, which do have word vectors, so it seems that this short phrase could easily be expressed in the word vectors semantic space; word vectors for collocations can easily be produced with a slight change in the front-end of the training algorithms, as mentioned in Le and Mikolov [15]. Mitchell and Lapata [21], who considered nine different $f_i : S \times S \rightarrow S$ as candidate functions for (two-word) phrase composition, under the hypothesis that the result of the composition should be a vector in the same semantic space as the original two words; and Coecke et al. [22], who proposed a project to build an algebra for phrase composition in the word semantic space, and offered a mathematical foundation for it.

Mapping phrase meanings in word semantic space is consistent with the information-retrieval origins of vector meaning. It has the intuitive problem that phrases, sentences, paragraphs, and documents have indefinitely greater range of meaning than single words. On the other hand, given the possible information content of 300-element vectors of 64-bit real numbers, it is plausible that we could easily encode almost all sentences in so many bits.

Mitchell and Lapata [21] found that the phrase-compounding functions for which the cosine distances best correlated with human judgements for phrase similarity were commutative, that is $f(x, y) = f(y, x)$. Still they consider it counter-intuitive word-order should not be a factor in composition. *dog bites man* does not convey the same information as *man bites dog*.

Many researchers have argued that analogy is the core of cognition and have tried to address different aspects of meaning by solving word analogy problems [23–25]. The intrinsic evaluation introduced by [8] has received much attention in recent years. For example, the analogy

$$\text{king} : \text{queen} :: \text{man} : \text{woman}$$

estimated by the vector equation

$$\text{king} - \text{queen} \approx \text{man} - \text{woman}$$

suggests that word vectors encode information about gender. By designing appropriate analogy questions, we can implicitly test different semantic and syntactic properties of semantic spaces. Several authors have mentioned weaknesses of word analogy evaluation. Linzen [26] showed that in some cases the solution is simply a nearest neighbor to the third word in the analogy question. Drozd et al. [27] studied various retrieval methods in addition to vector differences to solve analogy questions using the information contained in the semantic space, and mentioned inconsistency in results. Despite these weaknesses, word analogies are still one of the most commonly used intrinsic evaluation schemes.

In this paper, we consider the difference-vectors central to word analogy computations as building blocks for phrases, and word analogies as typed operators. They are typed because we do not expect, e.g. the plural operators for nouns and verbs to have the same geometric results in the semantic space, or the past-tense operator for infinitive verbs to provide a meaningful result for adjectives.

2 Methods

2.1 Arabic morphology and orthography

An introduction to Arabic morphology and orthography will be helpful in understanding how some Arabic phrases can be a single word. In computational linguistics, *word* has several senses.

- An *orthographic word* is characters delimited by spaces or punctuation. In addition to the other words in this sentence, *Don't* and *1984* are examples of orthographic words in English. Arabic examples include words with object or possessive suffixes, like *she will shave him* ستحلقه or *her house* بيتها.
- A *headword*, or *lemma*, is a dictionary entry. In English, the headword for a verb is the infinitive form, for example the word *walk*. In Arabic, it is the third-person singular perfect, for example the headword *he walked* مشى.
- A *conjugated form* is a headword marked for a limited set of features, including tense and number for verbs, case and number for nouns. In English *walked* and *walking* are conjugated forms of the headword verb *walk*. *Walks* is either the plural of the noun *walk* or the third person singular present of the verb. In Arabic, *he walks* يمشي and *she walks* تمشي are conjugated forms of the headword *he walked* مشى.
- A *derived word* is a word formed from other words. The English word *walker* is an example of a new headword formed from *walk*. The rule for deriving it is so well-known that it might not get an entry in every dictionary, but it is a new headword, a noun formed from a verb. Similarly, the Arabic word *pedestrian* is derived from the headword *he walked* مشى.

Arabic has both derivational and conjugational morphology. Derivational morphology describes the processes by which new headwords can be built from existing ones. In English, we can often add the suffixes *-er* or *-ing* to the end of a verb to make new nouns. For example, the English verb *write* can produce the nouns *writer* or *writing*. Similarly, the Arabic headword *he wrote* كتب can produce the nouns *clerk* كاتب or *book* كاتب. Unlike English, most derivational morphology in Arabic involves internal changes in the stem of the lemma, primarily lengthening or other changes in vowels, but also inserting consonants according to a library of fixed patterns. Arabic linguists call this *root and pattern morphology*. We do not deal with Arabic derivational morphology in any of our analogies. Even though there are semantic regularities, many words were formed long ago and have a history which has modified their meaning, just as *information* in English is only obscurely related in meaning to *form*, from which it seems to be derived by simple, regular affixes and in Arabic, *battalion* كتيبة, seems only remotely related to the headword *he wrote* كتب from which it comes.

Conjugational morphology describes the changes made to a noun to change its number (singular, dual, or plural) or its definiteness, or its case (nominative, accusative, or genitive); to an adjective to agree with the noun it modifies for definiteness, gender, number, and case; to a verb for tense (perfect or imperfect) and voice (active or passive), or in order to agree with its subject in person, number, and gender. Several of our analogies are based on conjugational morphology.

Arabic conjugational morphology uses both root and pattern morphology, and affixes. However, the only word vector space development algorithm we consider in this paper which is in any way aware of character strings within an orthographic word is `fastText` [10]. All the others consider a word as an indivisible unit, delimited by space or punctuation, usually denoted internally by an integer index.

For this reason, Arabic orthographic conventions matter. As in English, conjugation affixes are written without spaces to separate them from the stem, including (for nouns) the definite article *the* ال, and the possessive pronouns, and for verbs and prepositions, object pronouns. Also, the one-letter conjunctions *and* و and *then* ف; and the one-letter prepositions *in* ب , *for* ل, and *like* ك do not stand alone, but are attached to the following word. So for a noun we could have:

{*conjunction*} {*preposition*} {*article*} *noun* {*possessive pronoun*}

and for a verb:

{*conjunction*} *conjugated-verb* {*object pronoun*}

where the items inside {} may or may not occur.

Arabic pronoun subjects may be dropped, since the person, gender, and number are clear from the conjugated verb. So a single connected word with a verb could be an entire independent clause, with an implied subject, a verb, and a pronoun object. Similarly, a single connected word could be a prepositional phrase, functioning as an adverb or an adjective.

In this study, we take advantage of this property of Arabic orthography by determining to what extent the meaning of (some simple) phrases is the result of composition. The phrases we consider are those which the word-segmentation algorithms treat as a single word. We ignore other "phrases" which we consider errors. For example, there are Unicode codepoints for Arabic punctuation characters, but some word-segmentation algorithms treat them as regular characters, resulting in "words" containing punctuation, usually a trailing ،.

A final property of Arabic orthography is that short vowels are seldom written, so that many short words (and a few longer ones) are homographs, that is, have two distinct readings. For example, the Arabic written word علم could be the word *flag* عَلَم, the word *science* عِلم, or the word *teach* عَلَّم. The Arabic reader must decide the pronunciation based on the surrounding context. The word *read* is an example of the same phenomenon in English; it is pronounced differently in the sentences *I will read you a story*, in which *read* is an infinitive verb, and rhymes with the word *reed*; and *I read a good book last night* in which *read* is the past tense form, and is pronounced to rhyme with the word *red*.

2.2 Building the corpus

Word analogies have become a standard method for evaluating the quality of a semantic space. Mikolov's original example is that the meaning of word *king* is related to *queen* in the same sense as *man* is related to *woman*. The vector corresponding to *king* is subtracted, using ordinary vector arithmetic, from the vector for *queen*, and that difference is added to the vector for *man*. We search the neighborhood around the result vector, and find that the nearest word to that spot is the vector for *woman*. Apparently, the difference-vector incorporates the gender information, at least for some nouns.

To evaluate semantic spaces, we use files (or *relations*) of similar pairs, and any two different pairs from the same relation form an analogy which can be tested. Using a standard database of analogies, we can compare various algorithms for creating semantic spaces. Using translations of Mikolov's sets of English analogies, we can judge how well our algorithms and corpora succeed in building useful semantic spaces for other languages. Elrazzaz et al. [28] have prepared a corpus of Arabic word analogies, similar to Mikolov's English analogies, but with a focus on Arabic categories and morphology.

In this paper, we set out to do something a little different. Because the conventions for written Arabic mean that thoughts which would require several words to express in English can be written in a single Arabic word, we wanted to use Arabic to explore the semantic relationships between phrases and their parts. In particular, we look at vector addition as a compounding operation. The success rate to a particular relation shows whether the semantic space encodes this relation geometrically. This is interesting for other languages as well. For example, are verb tense changes adequately expressed by analogies, or do different semantic classes of verbs express tense differently? Our analogies examine a handful of morphological, syntactic, and semantic correspondences.

We created a corpus[1] consisting of 27 analogy types and approximately 20 pairs for each. Examples are shown in Table 1. Each analogy type corresponds to a Modern Standard Arabic morphological variation. For example, the *iswas* type consists of pairs of masculine singular imperfect and perfect verbs (roughly corresponding to present and past tense in English). The *noun-noun-h* type consists of pairs of a noun and the same noun with a masculine singular possessive suffix pronoun. The *verb-verb-hA* type consists of pairs of masculine singular imperfect verbs, with and without a feminine object pronoun. The *verb-she-hes* type consists of pairs of masculine and feminine singular imperfect verbs. The *noun-mA* consists of nouns and their masculine accusative case forms. Since the noun suffix اً denoting the masculine indefinite accusative case in classical Arabic is the only case ending which consistently appears in standard orthography, we thought it would be interesting to see whether the contextual algorithms which are used to build semantic spaces would recognize this case relationship.

We chose the sets of pairs by using regular expressions to search through the most common words in the Arabic Wikipedia. We rejected pairs in which one member of the pair was much less common than the other.

For the *vshe.../vhe...* series of fifteen analogies, we also rejected pairs in which the imperfect feminine form was a homograph for a perfect masculine form, as in the example of *she carries* تَحْمِل and *he endured* تَحَمَّلَ as described in section 2.5 below. So *verb-she-hes* is quite similar to *vhe-vshe*, except that the pairs are reversed, and there are no homograph problems of this particular kind, and the relations: *verb-verbh* and *vhe-vhehim*; *verb-verbhA* and *vhe-vheher*; are likewise similar except for that homograph check.

Twenty-four (twenty-seven counting all six relations just mentioned) analogy types is nowhere near enough to test the whole range of Arabic morphology, but it is enough to explore some of the problems.

2.3 Various Arabic semantic spaces

The backbone principle of methods for representing the meaning in a semantic space is the *Distributional Hypothe-*

1 Our corpus is available on the web at http://computersystemsartists.net/RelationCorpus.zip

Table 1: Our analogy types with sample pair, translation to English, and the number of word pairs.

Analogy	Example in Arabic	Translation to English	Pairs
iswas	تبع vs. يتبع	*he follows* vs. *he followed*	20
noun-b-noun	بإضافة vs. إضافة	*addition* vs. *in addition*	20
noun-bil-noun	بالبيت vs. بيت	*house* vs. *in the house*	20
noun-definite	الاشتراك vs. اشتراك	*participation* vs. *the participation*	29
noun-hA-noun-h	عمرها vs. عمره	*her age* vs. *his age*	20
noun-mA	تقريباً vs. تقريب	*approximation* vs. *approximately*	20
noun-noun-h	مسيرته vs. مسيرة	*parade* vs. *his parade*	20
noun-noun-ha	مسيرتها vs. مسيرة	*parade* vs. *her parade*	20
noun-w-noun	وروابط vs. روابط	*connections* vs. *and connections*	21
verb-she-hes	تقدر vs. يقدر	*he is able* vs. *she is able*	33
verb-verb-h	يجعله vs. يجعل	*he makes* vs. *he makes it* (masc. object)	18
verb-verb-hA	يجعلها vs. يجعل	*he makes* vs. *he makes it* (fem. object)	19
vheher-vhehim	يعتبره vs. يعتبرها	*he considers it* (fem. obj) vs. *he considers it* (masc. obj)	28
vheher-vshe	تعتبر vs. يعتبرها	*he considers it* (fem. obj) vs. *she considers*	23
vheher-vsheher	تعتبرها vs. يعتبرها	*he considers it* (fem. obj) vs. *she considers it* (fem. obj)	25
vheher-vshehim	تعتبره vs. يعتبرها	*he considers it* (fem. obj) vs. *she considers it* (masc. obj)	23
vhehim-vshe	تعتبر vs. يعتبره	*he considers it* (masc. obj) vs. *she considers*	28
vhehim-vsheher	تعتبرها vs. يعتبره	*he considers it* (masc. obj) vs. *she considers it* (fem. obj)	33
vhehim-vshehim	تعتبره vs. يعتبره	*he considers it* (masc. obj) vs. *she considers it* (masc. obj)	25
vhe-vheher	يعتبرها vs. يعتبر	*he considers* vs. *he considers it* (fem. obj)	25
vhe-vhehim	يعتبره vs. يعتبر	*he considers* vs. *he considers it* (masc. obj)	23
vhe-vshe	تعتبر vs. يعتبر	*he considers* vs. *she considers*	28
vhe-vsheher	تعتبرها vs. يعتبر	*he considers* vs. *she considers it* (fem. obj)	26
vhe-vshehim	تعتبره vs. يعتبر	*he considers* vs. *she considers it* (masc. obj)	23
vsheher-vshehim	تعتبره vs. تعتبرها	*she considers it* (fem. obj) vs. *she considers it* (masc. obj)	28
vshe-vsheher	تعتبرها vs. تعتبر	*she considers* vs. *she considers it* (fem. obj)	23
vshe-vshehim	تعتبره vs. تعتبر	*she considers* vs. *she considers it* (masc. obj)	23

sis [5], which states that the word meaning is related to the context where it usually occurs. Thus it is possible to compare the meanings of two words by statistical comparisons of their contexts.

Let $w \in \boldsymbol{V}$ denote a word, where $\boldsymbol{V}$ is a vocabulary of a language. A semantic space is a function $S : \boldsymbol{V} \mapsto \mathbb{R}^d$ which projects w into Euclidean space with dimension d. The meaning of the word w is represented as a real-valued vector $S(w)$.

In this work, we experiment with semantic spaces built with four different architectures.

- **CBOW** (Continuous Bag-of-Words) [8] is a simple neural network which tries to predict the current word according to the small context window around the word.
- **SkipGram** (SG) is an architecture similar to CBOW, but instead of predicting the current word based on the context, it predicts the context based on the word [8]. Both models CBOW and SkipGram are often denoted as *word2vec*.
- **GloVe** (Global Vectors) [9] model focuses more on the global word distribution in the data. GloVe is a log-bilinear regression model which employs a weighted least squares method to estimate the word vector representations.
- **fastText** is an extension of SkipGram model, which represents the word as a bag of character n-grams [10].

Table 2: Semantic spaces used for experiments.

Name	Architecture	Dim.	Data source	Data size
wiki-CBOW [11]	CBOW	300	Wikipedia	78.9M
wiki-SG [11]	SkipGram	300	Wikipedia	78.9M
wiki-fastText [8]	fastText	300	Wikipedia	78.9M
tw-CBOW [11]	CBOW	300	Twitter	1.1B
tw-SG [11]	SkipGram	300	Twitter	1.1B
web-CBOW [11]	CBOW	300	Web	2.2B
web-SG [11]	SkipGram	300	Web	2.2B
var-CBOW [12]	CBOW	300	Various	5.8B
var-SkipGram [12]	SkipGram	300	Various	5.8B
var-GloVe [12]	GloVe	300	Various	5.8B

Table 2 shows the settings of semantic spaces we use for experiments. For all models we use word vectors available on the web, and pre-trained on different types of data. CBOW and SkipGram models trained on Arabic Wikipedia, Twitter, and Web-crawled data were built by Soliman et al. [11][2]. Zahran et al. [12] collected large amount of raw Arabic texts from various sources[3] and trained CBOW, SkipGram, and GloVe. For the `fastText` model we use vectors which were pre-trained on Arabic Wikipedia[4].

The Arabic Wikipedia and the newspaper articles found in the Arabic Gigaword corpus are formal, deliberately non-dialectical Arabic. Twitter data often includes some dialect, but a significant fraction is formal.

For all semantic spaces we apply two post-processing techniques. First, we move the space towards zero (column-wise mean centering) and second, we normalize word vectors to be unit vectors.

The word analogy task consists of questions of the form: word w_1 is to w_2 as word w_3 is to w_4, where the goal is to predict w_4. We follow the definition from [8] and represent the word analogies according to vector offsets. To find the word w_4 (related to w_3 in the same way as w_2 is related to w_1), we go through all words w in vocabulary $\boldsymbol{V}$ looking for the word most similar to $S(w_2) - S(w_1) + S(w_3)$ according to cosine similarity. In our case, all word vectors are unit vectors so that the final equation has the following simple form

$$\hat{w}_4 = \arg\max_{w \in \boldsymbol{V}} \left(S(w) \cdot \left(S(w_2) - S(w_1) + S(w_3) \right) \right) \quad (1)$$

The input question words (i.e., w_1, w_2, and w_3) are discarded during the search as recommended by [8]. Finally, if $\hat{w}_4 = w_4$, we consider the question to have been answered correctly.

2.4 Comparing the semantic spaces using our corpus

We process the questions and calculate the accuracy as defined in subsection 2.3. During the search for an answer we always browse the 150,000 most frequent words in the corresponding dataset. We calculate the accuracy for each analogy type separately. Our word analogy corpus consists of common words in Arabic so that there are no out-of-vocabulary words for any tested semantic spaces. For each analogy type we process all combinations of pairs, but we omit the questions composed from two same pairs (e.g. for the category *noun-definite*, we have $29 \times 28 = 812$ questions).

Table 3: Average accuracies across all analogy types.

	acc@1	acc@5	acc@10
wiki-CBOW	34.3	50.5	62.8
wiki-SG	29.2	48.8	64.6
wiki-fastText	30.2	54.1	74.7
tw-CBOW	26.7	42.5	55.5
tw-SG	19.1	35.9	50.4
web-CBOW	40.3	58.6	71.3
web-SG	28.4	47.3	61.9
var-CBOW	**46.7**	**67.6**	**81.4**
var-SG	38.4	63.1	79.9
var-GloVe	42.7	65.2	81.0

Global results are shown in Table 3. For each semantic space, the final accuracy is always an average over accuracies for individual categories. This is motivated by the fact that for each analogy type, we have a different number of word pairs (see Table 1). By averaging the accuracies, each analogy type contributes equally to the final score. Acc@1 denotes the accuracy considering only the most similar word as a correct answer. Acc@5 and Acc@10 assume that the correct answer is in the list of five and ten most similar words, respectively. All accuracies are expressed in percentages. Overall, the best performance is achieved by CBOW trained on nearly six billion tokens from various sources. Also, CBOW architecture seems to be most suitable for Arabic analogies as it works best among all data sources.

2 Available at https://github.com/bakrianoo/aravec

3 Texts include Wikipedia, Arabic Gigaword Corpus, OpenSubtitles, etc. Pre-trained models are available at https://sites.google.com/site/mohazahran/data

4 fastText vectors for many languages trained on Wikipedia are available to download at https://fasttext.cc

Table 4: Accuracies (acc@1) for each individual analogy type and for each semantic space.

	wiki-CBOW	*wiki-SG*	*wiki-fastText*	*tw-CBOW*	*tw-SG*	*web-CBOW*	*web-SG*	*var-CBOW*	*var-SG*	*var-GloVe*
iswas	**41.3**	28.2	15.8	25.3	6.1	30.0	10.5	31.3	25.5	29.7
noun-b-noun	26.3	14.2	18.7	30.5	15.3	28.7	18.2	**36.6**	20.5	23.7
noun-bil-noun	10.8	**15.0**	12.4	2.4	3.9	3.9	1.8	14.5	11.3	13.4
noun-definite	23.9	15.3	11.8	30.2	19.8	21.6	9.4	30.3	23.4	**35.3**
noun-hA-noun-h	63.9	52.9	51.3	53.7	41.3	69.2	50.0	75.0	61.1	**75.3**
noun-mA	2.7	0.9	5.5	**6.4**	3.6	0.0	0.9	1.8	3.6	**6.4**
noun-noun-h	**49.5**	34.7	24.2	34.2	17.1	47.4	23.4	44.5	27.6	30.0
noun-noun-ha	**39.5**	29.7	17.6	26.6	14.7	34.7	13.9	28.2	14.5	23.4
noun-w-noun	39.3	51.4	47.6	44.3	48.1	61.4	40.5	71.4	75.0	**77.6**
verb-she-hes	72.5	64.2	51.4	75.6	64.5	71.1	45.0	**77.8**	69.0	75.6
verb-verb-h	38.6	23.2	39.2	16.0	12.1	**47.4**	35.0	38.6	25.5	32.0
verb-verb-hA	39.8	31.3	32.2	22.5	12.3	**49.7**	35.4	35.7	28.9	28.4
vheher-vhehim	53.8	45.9	75.9	29.5	30.6	66.3	73.7	**85.3**	85.1	82.1
vheher-vshe	23.7	27.3	2.8	25.9	19.8	31.6	8.9	36.2	22.3	**42.9**
vheher-vsheher	39.0	39.3	61.0	33.0	28.2	70.7	63.0	**92.7**	91.7	83.7
vheher-vshehim	16.4	21.9	3.0	9.7	7.1	24.9	15.4	**35.4**	24.1	32.2
vhehim-vshe	12.7	18.7	2.4	24.9	16.4	25.9	11.8	41.0	29.2	**43.5**
vhehim-vsheher	21.4	22.7	1.8	6.4	4.5	31.3	28.2	**45.3**	34.9	34.6
vhehim-vshehim	46.5	44.2	46.8	28.3	17.8	61.5	46.3	**77.3**	74.8	72.8
vhe-vheher	36.2	15.7	29.7	25.7	8.7	40.0	30.2	**40.7**	26.3	26.7
vhe-vhehim	34.8	23.3	40.7	20.4	16.0	**47.4**	35.8	41.3	30.0	35.4
vhe-vshe	**98.8**	93.4	84.7	91.5	74.2	88.2	58.3	97.8	92.7	97.8
vhe-vsheher	7.7	1.8	2.8	3.1	1.7	7.4	**13.5**	7.7	4.6	5.2
vhe-vshehim	12.1	3.4	1.2	5.3	1.2	10.9	9.5	**14.2**	5.3	11.7
vsheher-vshehim	25.9	30.4	62.7	15.1	15.1	40.1	34.4	**63.9**	57.1	63.2
vshe-vsheher	31.6	24.9	33.6	18.4	8.1	44.1	30.8	**47.6**	31.8	33.8
vshe-vshehim	18.2	14.2	38.7	15.0	6.7	33.4	23.3	**48.0**	39.9	35.4

Examining Table 4, the results for the *noun-noun-mA* relation are quite striking. Analogies between nouns and their indefinite-masculine-accusative forms seem not to work, which at first seems like a surprise. Considering the usage of the words in the list, however, removes some of the surprise. The so-called "Accusative forms" are mostly used as connecting forms like *as* وفقاً, *generally* عاماً, *approximately* تقريباً, *frequently* غالباً, *similarly* مثلاً, *previously* سابقاً, *exactly* تماماً, which are not used in any of the same contexts as their related nouns. We chose this set because they are quite frequent, and they are frequent because they are not topical, but used as general discourse markers. Uses of these expressions have no context in common with the related nominative forms.

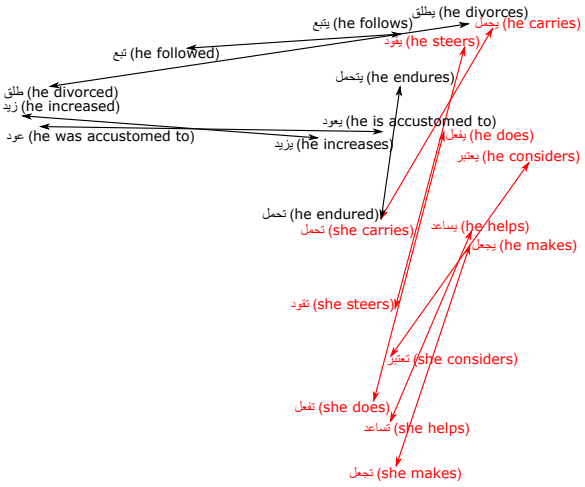
A similar effect seems to apply to expressions like *in spite of* بالرغم, *in fact* بالفعل, *completely* بالكامل, which are found in the *noun-bil-noun* relation. This relation scores second-lowest in Table 4, after *noun-mA*. Probably in both cases it would be possible to select words which are related in meaning to their nominative forms; but the pairs happen to be among the most common words in their form in the corpus.

Also striking is how scattered the best results are. For example, on the *iswas*, *noun-noun-h*, and *noun-noun-ha* relations, wiki-CBOW scored best, even though it is built on almost two orders of magnitude fewer words than the overall average best-scoring var-CBOW.

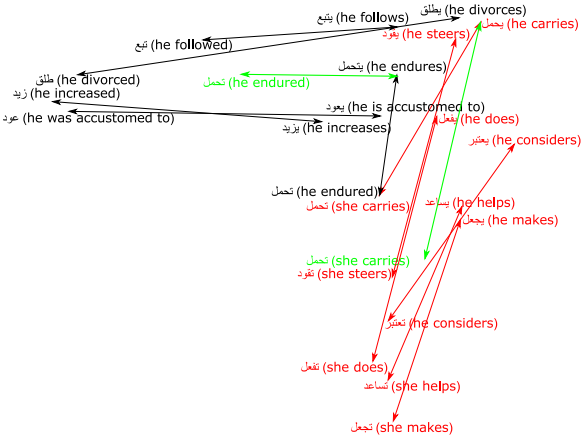
Table 5 and Table 6 show that when we look in a larger neighborhood for the correct answer, we are correspondingly more likely to find it.

2.5 Separating word senses

Figure 1, and Figure 2 are both visualizations of relations in the var-CBOW semantic space. Using Principal Component Analysis (PCA) we projected into two-dimensional space (parts of) relationships. Although the figures show nearly parallel vectors between members of the pairs, they also shows that the parallelism is not perfect.



(a) Unmodified positions



(b) Estimated positions for *she carries* تَحْمِل and *he endured* تَحَمَّلَ vectors in green

Figure 1: PCA projection of var-CBOW into two-dimensional space. Sample pairs for analogy *verb-she-hes* (e.g. *he does* vs. *she does*) are red and pairs for analogy *iswas* (e.g. *he does* vs. *he did*) are black.

One problem is that Arabic words frequently have homographs, distinct words which happen to be spelled the same way when vowels and other diacritics are omitted. For example, the words *she carries* تَحْمِل and *he endured* تَحَمَّلَ are both spelled تحمل when, as is usual, the vowels and diacritics are omitted.

In Figure 1, we see that the same spot is the terminus of two vectors, one from *he endures* يتحمل for the perfect form, and one from *he carries* يحمل for the imperfect feminine form. Neither vector is parallel to the others in its relation. We can estimate the proportion of each of the two meanings in the original corpus by looking at the relative frequencies of the two imperfect verbs, *he carries* يحمل rank 979 in frequency in the vocabulary and *he endures* يتحمل rank 14,639. Zipf's law lets us estimate that there is a factor of fifteen difference in the frequencies of the two, which might extend to the other conjugations of the two words.

In the figure, both vectors ending at تحمل are imperfectly parallel to the other difference-vectors in their corresponding relationships. If we were to separate the two words, and move the positions of the imaginary *she carries* تَحْمِل and *he endured* تَحَمَّلَ points (while maintaining the average position in its current spot) so that the difference-vectors are parallel to others in their respective relationships, the distance the imaginary *he endured* تَحَمَّلَ point would move is about fifteen times as much as the distance the *she carries* تَحْمِل point would move. That is, the location of تحمل in the semantic space is about where we would expect the frequency-weighted average of these two imaginary points to place it. This problem of homographs is an intrinsic feature of conventional Arabic orthography, so encountering it is no surprise. However, it happens to be worse in the var-CBOW semantic space, because Zahran et al. preprocessed their corpus to change ة to ه and terminal ي to ى. This is in accordance with Egyptian practice, although only a fraction of their sources were Egyptian. As a consequence, the word *queen* ملكة is a homograph of *his king* ملكه, and similarly for the majority of feminine nouns.

The homograph problem is more severe in Arabic than in English, but in both languages the *word sense* problem is significant. For example, a sense of *she carries* تحمل in both Arabic and English, which applies only to female subjects, is pregnancy. Fortunately for our exposition above, this sense was rare in our corpus.

Morphological word analogies can be used to disambiguate senses with different parts of speech, but semantic word analogies would be required to distinguish two senses with the same part of speech – an approach that requires significant supervision.

Table 5: Accuracies (acc@5) for each individual analogy type and for each semantic space.

	wiki-CBOW	*wiki-SG*	*wiki-fastText*	*tw-CBOW*	*tw-SG*	*web-CBOW*	*web-SG*	*var-CBOW*	*var-SG*	*var-GloVe*
iswas	**41.3**	28.2	15.8	25.3	6.1	30.0	10.5	31.3	25.5	29.7
noun-b-noun	26.3	14.2	18.7	30.5	15.3	28.7	18.2	**36.6**	20.5	23.7
noun-bil-noun	10.8	**15.0**	12.4	2.4	3.9	3.9	1.8	14.5	11.3	13.4
noun-definite	23.9	15.3	11.8	30.2	19.8	21.6	9.4	30.3	23.4	**35.3**
noun-hA-noun-h	63.9	52.9	51.3	53.7	41.3	69.2	50.0	75.0	61.1	**75.3**
noun-mA	2.7	0.9	5.5	**6.4**	3.6	0.0	0.9	1.8	3.6	**6.4**
noun-noun-h	**49.5**	34.7	24.2	34.2	17.1	47.4	23.4	44.5	27.6	30.0
noun-noun-ha	**39.5**	29.7	17.6	26.6	14.7	34.7	13.9	28.2	14.5	23.4
noun-w-noun	39.3	51.4	47.6	44.3	48.1	61.4	40.5	71.4	75.0	**77.6**
verb-she-hes	72.5	64.2	51.4	75.6	64.5	71.1	45.0	**77.8**	69.0	75.6
verb-verb-h	38.6	23.2	39.2	16.0	12.1	**47.4**	35.0	38.6	25.5	32.0
verb-verb-hA	39.8	31.3	32.2	22.5	12.3	**49.7**	35.4	35.7	28.9	28.4
vheher-vhehim	53.8	45.9	75.9	29.5	30.6	66.3	73.7	**85.3**	85.1	82.1
vheher-vshe	23.7	27.3	2.8	25.9	19.8	31.6	8.9	36.2	22.3	**42.9**
vheher-vsheher	39.0	39.3	61.0	33.0	28.2	70.7	63.0	**92.7**	91.7	83.7
vheher-vshehim	16.4	21.9	3.0	9.7	7.1	24.9	15.4	**35.4**	24.1	32.2
vhehim-vshe	12.7	18.7	2.4	24.9	16.4	25.9	11.8	41.0	29.2	**43.5**
vhehim-vsheher	21.4	22.7	1.8	6.4	4.5	31.3	28.2	**45.3**	34.9	34.6
vhehim-vshehim	46.5	44.2	46.8	28.3	17.8	61.5	46.3	**77.3**	74.8	72.8
vhe-vheher	36.2	15.7	29.7	25.7	8.7	40.0	30.2	**40.7**	26.3	26.7
vhe-vhehim	34.8	23.3	40.7	20.4	16.0	**47.4**	35.8	41.3	30.0	35.4
vhe-vshe	**98.8**	93.4	84.7	91.5	74.2	88.2	58.3	97.8	92.7	97.8
vhe-vsheher	7.7	1.8	2.8	3.1	1.7	7.4	**13.5**	7.7	4.6	5.2
vhe-vshehim	12.1	3.4	1.2	5.3	1.2	10.9	9.5	**14.2**	5.3	11.7
vsheher-vshehim	25.9	30.4	62.7	15.1	15.1	40.1	34.4	**63.9**	57.1	63.2
vshe-vsheher	31.6	24.9	33.6	18.4	8.1	44.1	30.8	**47.6**	31.8	33.8
vshe-vshehim	18.2	14.2	38.7	15.0	6.7	33.4	23.3	**48.0**	39.9	35.4

2.6 Parallel analogies

Are some of the analogies redundant? It might be the case that

he takes : he takes it :: she takes : she takes it

Figure 2 examines this for five of the twelve verbs which are in common for the four relations, *vshe-vshehim*, *vshehim-vhehim*, *vhe-vhehim*, and *vshe-vshehim*. Qualitatively, it looks like all of the blue lines are near parallel, and all of the red lines, likewise.

If the difference-vectors were independent of the starting point, the four endpoints of the four vectors would form a parallelogram, because the vector from *he does* to *she does* would be parallel to the vector from *he does it* to *she does it*. Similarly, the vector from *she does* to *she does it* would be parallel to the vector from *he does* to *he does it*.

Although this small set of examples shows significant variation from those expectations, all of the vectors start at the right and proceed to the left, and the deviations of the red vectors in the right half of the figure don't seem greater than the deviations of the red vectors in the left half of the figure.

One way to determine whether this qualitative observation is quantitatively correct would be to measure whether a relation made up by combining the two red relations is less parallel than either.

It is not, but the change in the variance is not what we would expect if combining two essentially identical relations.

Table 6: Accuracies (acc@10) for each individual analogy type and for each semantic space.

	wiki-CBOW	*wiki-SG*	*wiki-fastText*	*tw-CBOW*	*tw-SG*	*web-CBOW*	*web-SG*	*var-CBOW*	*var-SG*	*var-GloVe*
iswas	59.7	**60.8**	51.8	50.3	34.7	50.5	33.7	53.4	52.1	59.2
noun-b-noun	46.1	42.9	55.3	61.6	47.9	56.6	46.3	**64.2**	63.4	60
noun-bil-noun	24.7	30.3	36.1	16.6	23.9	26.1	17.1	48.4	**56.6**	35.8
noun-definite	54.9	55.4	65.1	65.1	58	61.1	38.2	75.1	79.1	**80.3**
noun-hA-noun-h	83.2	82.6	92.6	77.1	68.9	85.5	81.6	95.5	93.9	**98.2**
noun-mA	12.7	17.3	**37.3**	22.7	17.3	29.1	12.7	13.6	19.1	31.8
noun-noun-h	68.4	69.5	66.6	66.3	44.7	78.9	62.1	**84.5**	80.5	82.4
noun-noun-ha	69.7	70.5	67.1	68.4	54.7	72.6	51.1	76.1	71.8	**78.9**
noun-w-noun	78.3	90.2	93.8	80.5	84.3	92.1	91	98.1	99.3	**100**
verb-she-hes	87.7	89	82.2	84.3	86.5	82.3	75.1	**90.8**	86.9	87
verb-verb-h	82.7	78.8	**90.5**	49.3	50	81	70.9	85.9	87.9	86.9
verb-verb-hA	68.7	69	76.9	58.5	40.1	80.4	72.5	**86.8**	83.6	86
vheher-vhehim	79.2	80.2	92.9	58.1	60.6	91.7	98.9	99.7	99.3	**100**
vheher-vshe	57.5	61.1	82.4	59.5	49.6	74.9	47.6	**88.3**	82.6	85.6
vheher-vsheher	71.5	72	70	62.2	63.2	87.7	86	**100**	**100**	99.5
vheher-vshehim	47.4	56.7	77.7	35.8	33.8	58.1	45.8	**79.6**	75.9	74.3
vhehim-vshe	49.7	58.7	75.1	54.8	48.7	76.9	52.8	**90.5**	86.9	87
vhehim-vsheher	55.5	57	72.2	24.7	25.7	71.7	67.7	**90.7**	88.4	84.4
vhehim-vshehim	84.5	77.3	**99**	57	51.3	80.7	77.3	87.7	88.3	90.7
vhe-vheher	67.3	61.8	80.3	59.5	49.8	74	69.2	85.2	79.5	**85.8**
vhe-vhehim	68	75.5	**95.1**	62.1	62.8	80.4	76.9	87.7	86.2	88.9
vhe-vshe	**100**	**100**	**100**	97.4	93.8	99.7	91.1	99.9	99.7	99.9
vhe-vsheher	55.8	51.4	60.6	42.3	35.5	68.6	65.4	**82.5**	73.8	72.2
vhe-vshehim	52.2	45.1	49.2	39.1	30.2	47	36.8	**68.6**	63.8	65
vsheher-vshehim	56.2	65.1	86	43.9	46.3	66.5	66.5	85.4	**87.6**	87
vshe-vsheher	61.9	63.6	73.5	48.8	49.4	79.1	76.3	88.5	87.5	**89.9**
vshe-vshehim	51.8	61.1	87.4	52.2	48.6	72.5	60.1	90.9	82.8	**91.1**

Relation	Average angle from mean difference-vector	Std. dev.
vhe-vshe	44°	8.3
vhehim-vshehim	52 °	7.9
combined	51°	8.8

Similarly, we can measure the blue relations

Relation	Average angle from mean	Std. dev.
vshe-vshehim	57	5.7
vhe-vhehim	59°	6.1
combined	61°	5.8

Finally, what if we combine a red and a blue relation, say *vhe-vshe* and *vhe-vhehim*? We have the figures for the separate relations already, and the combined relation has an average deviation from the mean vector of 63° and a standard deviation of 12.1.

3 Discussion

We have built a corpus of simple morphological analogies in Arabic, which include semantics that would require several words and corresponding syntax in English. We have demonstrated that vector manipulation can be used to explain how morphological units can combine into phrases.

However, there are many possible confounding issues before phrase construction becomes simple arithmetic.

The most important problem for the difference vector method is the relatively large divergence from parallelism in difference vectors in similar analogies in the same relation. Because difference vectors are result of subtraction, and because the endpoints for morphological analogies tend to be near one another in the semantic space, small errors in placement caused by homographs, differ-

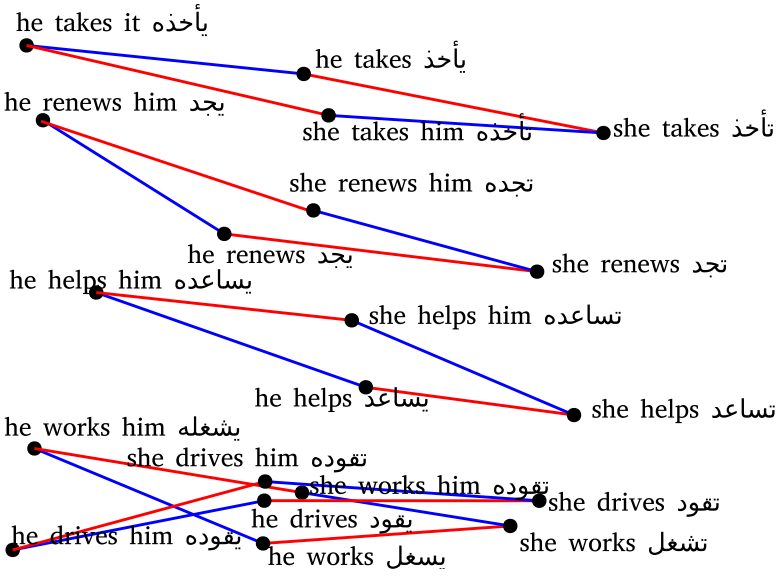


Figure 2: PCA projection of var-CBOW into two-dimensional space showing simple verb phrases: relations *vshe-vshehim* and *vshehim-vhehim* are red; relations *vhe-vhehim* and *vshe-vshehim* are blue. Each set of four verb phrases must be a quadrilateral when projected on a plane. Figure illustrates parallelism between opposite sides of quadrilateral.

ent sense mixtures, or low density of words in the corpus, can cause large divergences in the difference vectors. We see an average difference vector divergence (from the mean for the relation) ranging from 44° for *vhe-vshe* to 69° for *noun-definite* (these figures are for the corpus var-CBOW, which has the highest overall acc@1 accuracy on all analogies. The divergence angles are inversely correlated with the accuracy rate).

Another of these issues is the existence of set-phrases, or idioms, which have evolved a meaning separate from their constituents. The commonest phrases in the vocabulary are likely to have this property. For particular cases, we can discover that a phrase is non-compositional if the compositional word analogies don't work, that is if its semantic vector is distant from where the other word analogies of its type would place it. Ideally, if analogies are to be used to discover non-compositional phrases, the set of pairs in the relation should all be compositional. As we've already seen in the discussion of *noun-mA* in subsection 2.4, the current relations include non-compositional (idiomatic) pairs.

Arabic data poses a final challenge in that many words share the same written representation, and this influences both the position and nearest neighbors of words in the semantic space, in ways that depend on the frequency and usage of the various homographs.

One way to approach the homograph problem might be to prepare a corpus in which the vowels and other diacritics are all marked. Some of the tools for doing this, for example MADAMIRA [29] and the Farasa based tool [30] claim unsupervised error rates in the low single digits. In this case the errors introduced by the tool would be much less than those arising from homographic confusion.

However, given that Arabic words, even with the same pronunciation, can have several senses, solving the homograph problem may be worthwhile, but is not sufficient.

An issue which we have not considered is whether various kinds of normalization, which in general improve word analogy performance, might not have a different effect on morphological analogies, where the two members of a relation pair are usually close together in the semantic space before normalization.

All the issues of phrase composition are especially interesting in cross-lingual contexts, where a word in one language might be a phrase in another. A similar study could be made with, for example, an English semantic space which incorporates bigrams and trigrams in addition to single words. Data sparsity is likely to be a problem, but there is much more English data available.

References

[1] Turney P. D., Pantel P., From frequency to meaning: Vector space models of semantics, Journal of Artificial Intelligence Research, 2010, 37, 141–188

[2] Collobert R., Weston J., A unified architecture for natural language processing: Deep neural networks with multitask learning, In: Proceedings of the 25th International Conference on Ma-

chine Learning, 2008, 160–167
[3] Lample G., Ballesteros M., Subramanian S., Kawakami K., Dyer C., Neural architectures for named entity recognition, In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016), San Diego, California, Association for Computational Linguistics, June 2016, 260–270
[4] dos Santos C. N., Gatti M., Deep convolutional neural networks for sentiment analysis of short texts, In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, Dublin City University and Association for Computational Linguistics, August 2014, 69–78
[5] Harris Z., Distributional structure, Word, 1954, 10(23), 146–162
[6] Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., Harshman R., Indexing by latent semantic analysis, Journal of the American Society for Information Science, 1990, 41(6), 391–407
[7] Bengio Y., Ducharme R., Vincent P., Jauvin C., A neural probabilistic language model, Journal of Machine Learning Research, 2003, 3, 1137–1155
[8] Mikolov T., Chen K., Corrado G., Dean J., Efficient estimation of word representations in vector space, CoRR, 2013, abs/1301.3781
[9] Pennington J., Socher R., Manning C., Glove: Global vectors for word representation, In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, Association for Computational Linguistics, October 2014 , 1532–1543
[10] Bojanowski P., Grave E., Joulin A., Mikolov T., Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics, 2017, 5, 135–146
[11] Soliman A. B., Eissa K., El-Beltagy S. R., AraVec: A set of Arabic word embedding models for use in Arabic NLP, Procedia Computer Science, 2017, 117(Supplement C), 256–265
[12] Zahran M. A., Magooda A., Mahgoub A. Y., Raafat H., Rashwan M., Atyia A., Word representations in vector space and their applications for arabic, In: Gelbukh A. (Ed.), International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2015), Part I, LNCS 9041, Springer, 2015, 430–443
[13] Mikolov T., Yih W.-T., Zweig G., Linguistic regularities in continuous space word representations, In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013), 2013, 746–751
[14] Levy O., Goldberg Y., Linguistic regularities in sparse and explicit word representations, In: Proceedings of the Eighteenth Conference on Computational Language Learning, Association for Computational Linguistics, 2014, 171–180
[15] Le Q., Mikolov T., Distributed representations of sentences and documents, In: Proceedings of the 31st International Conference on Machine Learning, 2014
[16] Kiros R., Zhu Y., Salakhutdinov R. R., Zemel R., Torralba A., Urtasun R., Fidler S., Skip-thought vectors, In: Advances in neural information processing systems, 2015, 3294–3302
[17] Hill F., Cho K., Korhonen A., Learning distributed representations of sentences from unlabelled data, In: Proceedings of NAACL-HLT 2016, Association for Computational Linguistics, 2016, 1367–1377
[18] Wieting J., Bansal M., Gimpel K., Livescu K., Charagram: Embedding words and sentences via character n-grams, In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics , 2016, 1504–1515
[19] Sutskever I., Vinyals O., Le Q. V., Sequence to sequence learning with neural networks, In: Proceedings of the 28th Conference on Neural Information Processing Systems (NIPS 2014), 2014
[20] Conneau A., Kruszewski G., Barrault L., Lample G., Baroni M., What you can cram into a single vector: Probing sentence embeddings for linguistic properties, In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), 2018, 2126–2136
[21] Mitchell J., Lapata M., Composition in distributional models of semantics, Cognitive Science, 2010, 34, 1388–1429
[22] Coecke B., Sadrzadeh M., Clark S., Mathematical foundations for a compositional distributional model of meaning, Lambek Festschrit, special issue of Linguistic Analysis, 2010, 36, 345–384
[23] Turney P. D., Littman M. L., Bigham J., Shnayder V., Combining independent modules to solve multiple-choice synonym and analogy problems, In: Recent Advances in Natural Language Processing, 2003
[24] Turney P. D., A uniform approach to analogies, synonyms, antonyms, and associations, In: Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008), 2008, 905–912
[25] Jurgens D. A., Turney P. D., Mohammad S. M., Holyoak K. J., Semeval-2012 task 2: Measuring degrees of relational similarity, In: Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, SemEval '12, Stroudsburg, PA, USA, Association for Computational Linguistics, 2012, 356–364
[26] Linzen T., Issues in evaluating semantic spaces using word analogies, In: Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP, 2016, 13–18
[27] Drozd A., Gladkova A., Matsuoka S., Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen, In: Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016): Technical Papers, 2016, 3519–3530
[28] Elrazzaz M., Elbassuoni S., Helwe C., Shaban K., Methodical evaluation of Arabic word embeddings, In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), July 2017, 454–458
[29] Pasha A., Al-Badrashiny M., Diab M., Kholy A. E., Eskander R., Habash N., et al., MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic, In: Proceedings of the 9th International Conference on Language Resources and Evaluation, 2014, 1094–1099
[30] Darwish K., Mubarak H., Abdelali A., Arabic diacritization: Stats, rules, and hacks, In: Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP), 2017, 9–17