Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 142 (2018) 83–93

Procedia Computer Science

www.elsevier.com/locate/procedia

The 4th International Conference on Arabic Computational Linguistics (ACLing 2018),
November 17-19 2018, Dubai, United Arab Emirates

Morphological Word Embedding for Arabic

Rana Aref Salama[a], Abdou Youssef[b], Aly Fahmy[a]

[a]Cairo University, Giza City, Cairo 12613, Egypt
[b]George Washington University, Washington, DC 20052 , USA

Abstract

Word embedding has opened new and exciting avenues for understanding and processing languages. The simple yet effective word embedding models rapidly became a dominant building block for Natural Language Processing(NLP) applications as they impressively encode linguistic similarities and syntactic regularities between words. However, ignoring the morphological structure of words degrades its performance when applied to languages with complex morphology like Arabic. In this paper, we investigate enhancing Arabic word embedding by incorporating morphological annotations to the embedding model. We further tune the generated word vectors to their lemma forms using linear compositionality to generate lemma-based embedding. To assess the effectiveness of our model, we perform evaluation using Arabic analogy, sentiment and subjectivity analysis. Our results show improvements over existing state-of-the-art methods for Arabic word embedding.

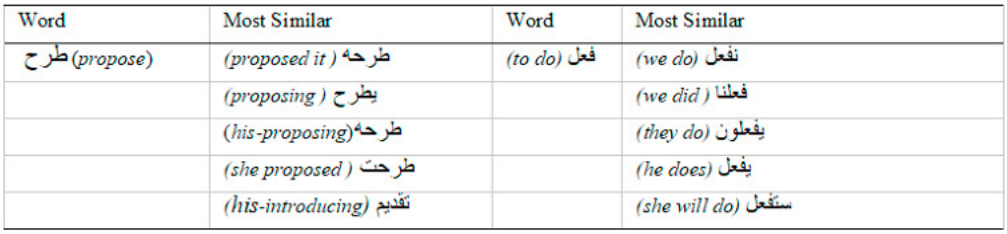
Keywords: Word Embedding; Word2Vec; Arabic; Morphology; NLP

1. Introduction

A point in a two-dimensional space is a primitive notion for a unique position set by two values, usually x and y. Similarly, word embeddings are unique vectors of features influenced by all other words in a distributional space. While dictionaries describe meaning of words through linking concepts to other concepts, word embedding link words to numbers rather than words to encode word-level semantics and similarities [21]. Additionally, word embedding can capture many linguistic regularities and patterns [24]. However, they treat words holistically with no attention to their internal structure, ignoring the fact that, morphologically, meaning is a multi-faceted concept with multiple axes along which two words can be similar [12]. For languages known to be morphologically impoverished, like English, this fact could be ignored but for processing inflectional languages with rich morphology, like Arabic, exploiting word internal structure is mandatory [13]. For example, applying word embedding to Arabic causes different surface realizations of a word, like conjugations and inflections of a verb, to be the most semantically similar words and this strongly demotes other semantically similar words that have different forms. In addition, these kinds of similarities lead to

[*] Email: r.aref@fci-cu.edu.eg, ayoussef@gwu.edu, aly.fahmy@cu.edu.eg

10.1016/j.procs.2018.10.463



Word	Most Similar	Word	Most Similar
طرح *(propose)*	*(proposed it)* طرحه	*(to do)* فعل	*(we do)* نفعل
	(proposing) يطرح		*(we did)* فعلنا
	*(his-proposing)*طرحه		*(they do)* يفعلون
	(she proposed) طرحت		*(he does)* يفعل
	(his-introducing) تقديم		*(she will do)* ستفعل

Table 1. Examples of Semantic Similarity for Arabic Words.

sparsity problems and deficiency in exploiting shared semantics [13]. As shown in Table 1, the most similar words to the word طرح (i.e., propose) are all variants of its base form except the word تقديمه which means introducing or giving. Similarly, the most similar words to فعل (i.e., to do) are all variations of the same word.

Accordingly, enriching word embeddings with substantial grammatical information can generate more meaningful embeddings that are capable of capturing both semantic and morphological similarities as affected by different types of linguistic properties [5].

Therefore, in this paper, we opt to study the effect of incorporating morphological information to Arabic word embedding in a way that enhances both semantic and morphological similarity. We present two models; morphological embedding that incorporates POS tags with words for embedding, and lemma-based embedding that represents morphologically related words using their lemma form. We integrate recent adavnces in Arabic NLP in morphological analysis [33, 26], benchmark and embedding evaluation [16]. We also incorporate recent advances in distributional semantic representation [25] to generate a multilevel approach for Arabic word embedding with particular focus on inflectional morphology. We base our work on the recent Arabic POS tagging and segmentation tool, Farasa[1], to generate morphological data relevant to our embedding.

The rest of this paper is organized as follows. In Section 2 we survey related work for Arabic and morphological word embedding. Our model is described in Section 3. Section 4 presents our evaluation results, and finally Section 5 concludes our work and summarizes proposed future work.

2. Related Work

Semantic representation using the distributional hypothesis has gained a resurgence of research efforts that spanned more than six decades. The basic idea of distributed semantics has a theoretical basis in structural linguistics and language philosophy since the 1950s. However, the early stages of semantic distribution models, like Latent Semantic Analysis(LSA), and automatic generation of contextual features started around the late 1990s. Figure 1 represents the most popular achievements in word embedding since that time [34]. As shown, distributed semantic representation evolved through time by refining early models and using advances in neural networks to predict a probability distribution for words given some surrounding words called context.

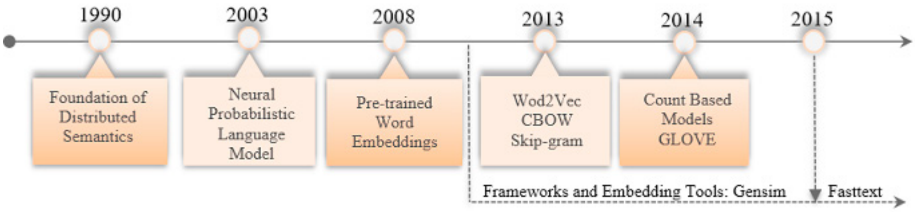

1990
2003
2008
2013
2014
2015
Foundation of Distributed Semantics
Neural Probabilistic Language Model
Pre-trained Word Embeddings
Wod2Vec CBOW Skip-gram
Count Based Models GLOVE
Frameworks and Embedding Tools: Gensim
Fasttext

Fig. 1. Distributed Semantics Evolution

[1] http://qatsdemo.cloudapp.net/farasa/

In 2003, Bengio et al. proposed a neural probabilistic language model using a shallow feed forward neural network that learned distributed representations for words [7]. Later, pre-trained word embeddings were proposed by Collobert et al. in 2008 [11]. In 2013, Mikolov et al. proposed the booming word2vec model with its two well-known architectures: Continuous Bag of Words (CBOW) and Skip-gram models. Key features of word2vec popularity include its simple structure, reduced computational complexity, and its ability to build low-dimensional vector representations that effectively capture semantic similarities and syntactic regularities [19, 34, 25]. Subsequently, models like Glove [27] integrated proximity and global context to capture multi-dimensional word semantics. Consequently, implementation of these models became available through different libraries and packages like Gensim [28] and Fasttext [18].

2.1. Arabic Word Embedding

Arabic word embedding serves as the basis for many recent NLP tasks for Arabic. They have been effectively used in sentiment and subjectivity analysis [15, 2, 4], neural machine translations [29],word sense disambiguation[20], textual entailment [3] and semantic similarity[26]. However, the performance of these tasks directly depends on the quality of word embedding, which in return strongly depends on the implemented model and used corpus [21]. The best models known for Arabic word representations are CBOW and Skip-gram(SG) models [16]. As shown in Figure 2, Skip-gram embeds a word by predicting its surrounding words, while CBOW embeds a word by predicting the word itself given its surroundings. Both models are considered simple, computationally efficient, and suitable for large datasets [27]. On the other hand, the nature of Arabic as a morphologically rich language with high character variation has its direct impact on how influential a corpus is for generating good embeddings. Additionally, the nature of the corpus, as to whether it is in dialectal or standard Arabic, affects how well should a corpus be cleaned and normalized to guarantee a richer semantic representation[33].

2.2. Morphologically Based Embedding

One of the promising research directions for developing better word embeddings is enriching words with their morphological features to minimize frequent morphological variants [29]. In this section we briefly present how morphology was incorporated with word embedding models in languages other than Arabic, like English, at different levels. Some models leverage morphological information with words fed to the embedding module, like morphological POS tags [31]. Other models utilize the n-grams constitutents of words to implicitly exploit the meanings of morphemes (suffix, root and stem) before embedding [33]. Other models use a multi-task objective to encourage word embedding to reflect morphological tags [12]. Other methods are geared towards injecting morphological constraints to fine tune final vector representations to attract morphologically related words [32]. Other methods weight morphological relations during the embedding process [23, 29]. Some models create embeddings of n-gram constituents of a word rather than the whole word, where final word representations are the summation of these sub-representations [10, 8, 14, 9]. Morphological embedding has been investigated for other languages like Portuguese[17], German [12] and Swedish[6] with promising results, in addition to some achievements for morphologically rich languages like Turkish[13] and Hebrew[5]. To the best of our knowledge, there is no pretrained Arabic morphological word embeddings. However, just before publishing this paper, Pamela et al. proposed a modified Skip-gram model for Arabic word embeddings that concatenates words and lemmas using n-grams character structure for embedding [29].

3. Arabic Morphological Word Embedding

In this work we investigate incorporating morphological knowledge with Arabic word embedding to capture the two aspects of word similarity: semantic similarity and morphological similarity. We propose a model that leverages contextual information and morphological knowledge to generate two kinds of morphologically based embeddings. One is morphological embedding that incorporates linguistic properties, including POS tags, with words before embedding. The other is lemma-based embedding, where word lemmas are further used to perform morphological abstraction and semantic consolidation. Initially, we add a pre-processing phase to our model for generating morphological annotations using Farasa tagger. We then generate morphological embeddings and consequently apply unification and lemma fitting as described in subsection 3.3 to generate lemma-based embedding.

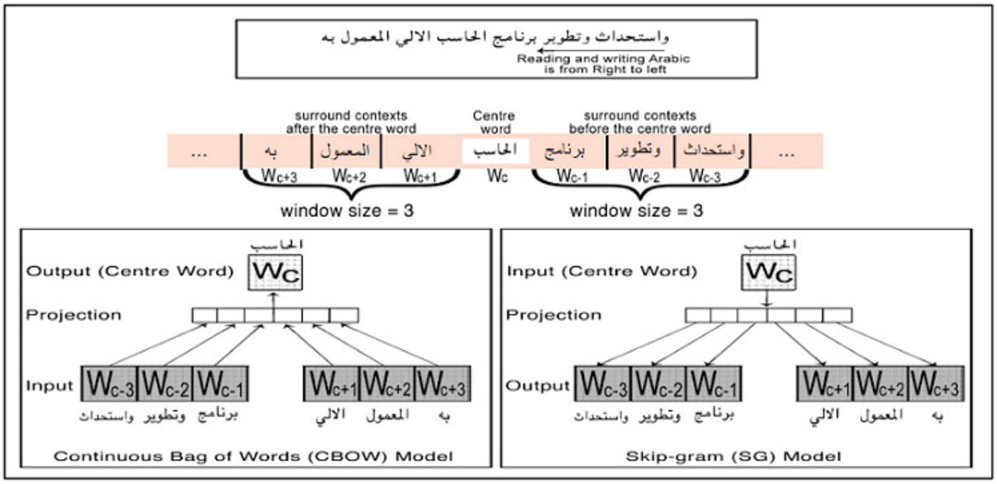

واستحداث وتطوير برنامج الحاسب الالي المعمول به
Reading and writing Arabic is from Right to left
surround contexts after the centre word
Centre word
surround contexts before the centre word
... به المعمول الالي الحاسب برنامج وتطوير واستحداث ...
Wc+3 Wc+2 Wc+1 Wc Wc-1 Wc-2 Wc-3
window size = 3
window size = 3
Output (Centre Word)
Projection
Input
Continuous Bag of Words (CBOW) Model
Input (Centre Word)
Projection
Output
Skip-gram (SG) Model

Fig. 2. Arabic Word2vec Model [2].

3.1. Pre-processing

Pre-processing tends to be an important step for word embedding to eliminate irrelevant tokens[22]. Although it is advised to be minimal in languages like English, yet for Arabic and morphological embedding, pre-processing is considered to be a crucial step to perform morphological annotation and enforce vocabulary reduction for better word representations. We formally undertake three phases of pre-processing as shown in Figure 3. Initially, the raw corpus is tokenized; then, cleaning, normalization and annotation is performed.

Cleaning consists of eliminating irrelevant, non Arabic words, stop words and punctuation characters. Any Arabic or English punctuation such as {? , ; ! { } ? # $ % }, special or foreign characters, are removed. We refer to the stopwords list proposed for Arabic in [1] to select a relevant list of 270 stopwords to be removed. Finally, some special tags are induced to maintain structure of data, like <SOS>to indicate start of sentence, <EOS>to indicate end of a sentence, <UNKNOWN>to indicate out of vocabulary words.

Normalization is concerned with characters mapping for linguistic reduction and standardization among the whole corpus as follows:

- Mapping characters with multiple forms into a single form. For example, the different forms of the ”aleph” letter (أ , إ , آ) are mapped to one form”ا”, and the form ”ta” letter ”ة” is mapped to ”ه”.
- Elongation removal, for example بناااات is mapped to بنات.
- Diacritics removal: the set of all arabic diacritics are to be mapped to Null.
- Mapping all digits, English and Arabic, to <NUM>tag.

Annotation is used for preparing data for morphological embedding. This implies tagging all Arabic words with their corresponding morphological categories. We apply cleaned and normalized data to Farasa for segmentation and annotation. The data is annotated with relevant POS tags to capture syntactic regularities.

3.2. Morphological Embedding Model

For morphological embedding, we follow the sense2vec model developed by Trask et al. [31]. The well-known word2vec model developed by Mikolov et al. [24, 25] is modified to generate a preliminary word vector representations for our model based on morphological annotation. We demonstrate our methodology using CBOW for its simplicity. Consider a sequence of words $S = [w_{i-c}, .., w_i, ..w_{i+c}]$ snapped from a training corpus D of size N using a fixed window of size c centered around w_i. Each word is coupled with a tag m from a corresponding set of morphologi-

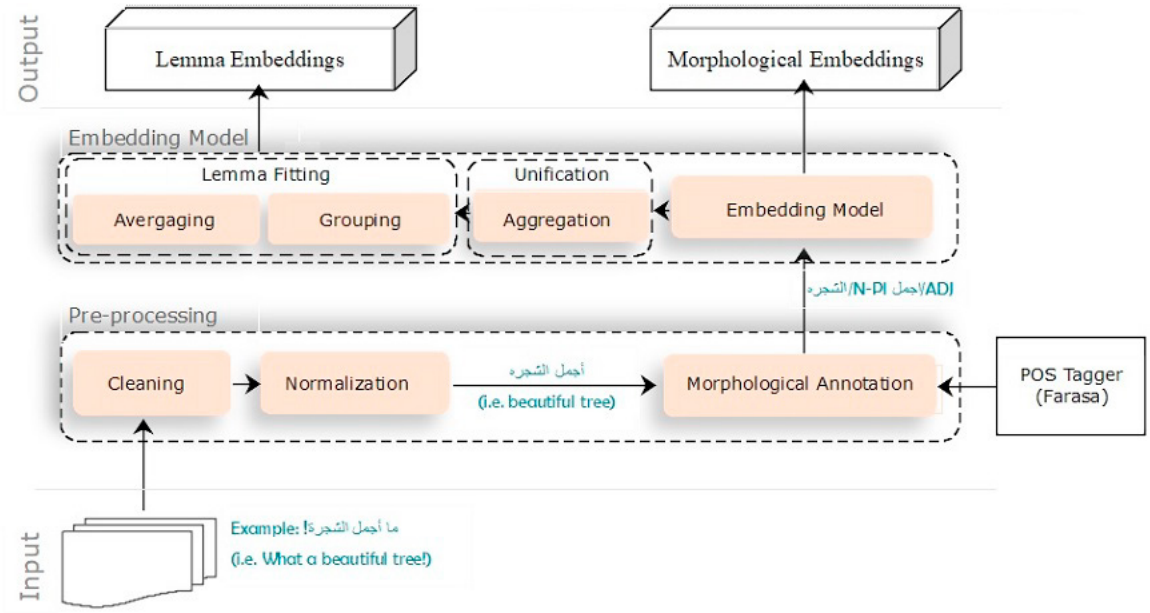


Fig. 3. Arabic Morphological Embedding Model

cal tags $M = m_1, ..., m_t$ acquired through pre-processing annotation. The learning objective function for log-likelihood maximization is updated as follows:

$$\mathcal{L}(D) = \frac{1}{N}\sum_{i=1}^{N} \log p(w_i, m|w_{cx}, m_{cx}) \tag{1}$$

where w_{cx} indicates the context of the current word w_i with tag m and m_{cx} indicates the corresponding morphological tag for every word in this context.

Similarly, the Skip-gram model can be updated to leverage morphological information and predicts surrounding words $(w_{i-c}, .., w_{i+c})$ given the current word w_i and their corresponding morphological tag.
Subsequently, we can directly use the generated morphologically embedded vectors to represent words given their morphological POS tag, and then apply different NLP tasks. For example, we can check the well known analogy question $V($ولد$, N) = V($رجل$, N) + V($امرأة$, N) - V($بنت$, N)$ (i.e., V(boy,N)=V(man,N)+V(woman,N)-V(girl,N)), noting that $V(w, m)$, indicates the vector of word w with tag m. Under this model, each word is accompanied by its morphological tag for processing. Accordingly, words have multiple embeddings with respect to their different senses and interpretations. This is considered one of the most important outcomes from morphological embeddings to handle multiple word senses and ambiguity.

3.3. Lemma-Based Embedding

Lemma-based embedding is another embedding schema where vectors encode more condensed semantic information regardless of their inflectional form[5]. To generate lemma-based embeddings, we initially unify all embeddings of different senses into a unified representation using the linear compositionality nature of word vectors [19].Unified representations encode related morphological embedding for a word and generate a single word embedding representation for this word. We aggregate all senses of a given word into a single word representation using vector addition. For example, we add the vectors (اكل$, N$) and (اكل$, V$) to generate a unified representation for the word اكل(i.e., eat). Formally, for a word w_n and a set of morphological tags $M = m_1, ..m_t$, the unified vector $V(w_n)$ is calculated as,

$$V(w_n) = \sum_{i=1}^{t} V(w_n, m_i) \tag{2}$$

Afterwards, the lemma-based embeddings are created through morphologically fitting the unified embedding space to corresponding word lemmas. We call this step lemma fitting. Lemma fitting is achieved by scanning the whole corpus for words that share the same lemma component and group them into a single group. Accordingly, words w_1, w_2 will be pulled to the same group of w_{lemma}, if $lemma(w_1)$=lemma(w_2). For example (عمل,بعمل,العمل), meaning (work, working, the work), will be in the group of the lemma عمل (i.e., work).
After grouping morphologically related words through lemma fitting, the final lemma representation is generated using the mean vector of all word vectors in a group as shown below, where G is the number of words in a single group and $V(w_{lemma})$ is the final vector representation for the lemma word.

$$V(w_{lemma}) = \frac{1}{G}\sum_{i=1}^{G} V(w_i) \tag{3}$$

4. Evaluation

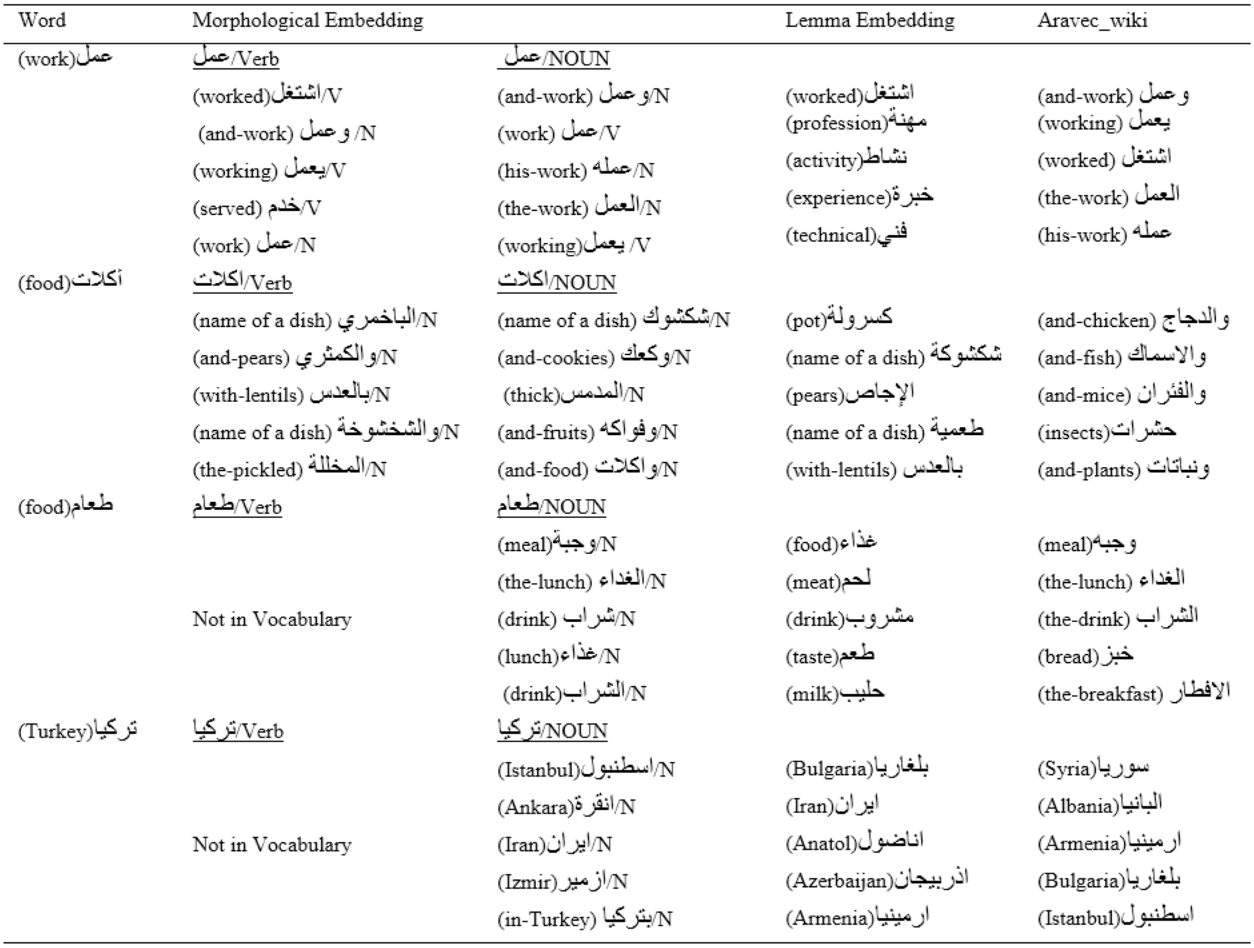
We evaluate our model using both intrinsic and extrinsic evaluation. We performed a subjective intrinsic evaluation by comparing our results with other Arabic word embedding models developed by Aravec [30] using an Arabic analogy benchmark [16]. We further apply extrinsic evaluation on Arabic sentiment and subjectivity analysis. We used the Arabic Wikipedia corpus to train our model using the word2vec model implemented in Gensim [28]. For simplicity and efficiency, we build our model using the CBOW technique. We found setting CBOW parameters to window of size 5 and vector of dimension 300 gives the best results. We initially used Farasa POS tagger and segmenter to annotate our data. We mainly focused on verb and noun POS tags for our evaluation.

4.1. Intrinsic Evaluation

We first tested our model using a random set of verbs and nouns to measure the quality of our word embeddings. Table 2 shows a snapshot of our results[2] as opposed to embeddings generated on the same Wikipedia corpus, with same parameters and model, using Aravec [30], we call it Aravec_wiki. We list the 5 nearest similar words for each word, using both morphological and lemma-based embeddings. We further evaluated our model using the recently developed Arabic word embedding evaluation methodology and its analogy benchmark [12]. The analogy benchmark is composed of more than 115,000 word analogy questions for Arabic with nine relations, each consisting of more than 100 word pairs for at least 100,000 tuples. Table 3 shows a sample analogy question for each relation. To measure our models' accuracy, we apply the same evaluation methodology used with the benchmark. We use the 2 versus 11 word pairs for our experiment with top-n words to consider a question answered correctly. Questions are wrongly answered if at least one of the words in the question is not found in the word embeddings, or the correct answer is not in the top-n predicted answers. Since CBOW model outperformed all other models in the benchmark evaluation, we use it as a basis for our model evaluation. In this benchmark, the authors based their evaluation on the embeddings generated by Zahran et al. using a very large corpus with 6.3 million vocabulary words [35], however, we couldn't use the same corpus for our embeddings due to its privacy. Given the fact that a model's accuracy is directly proportional to the size of a training corpus [21], it won't be fair to compare their results with ours since we used a corpus with around 200.000 vocabulary words. We used this benchmark for subjective evaluation with a similar model from Aravec that was proven to generate good word embeddings.

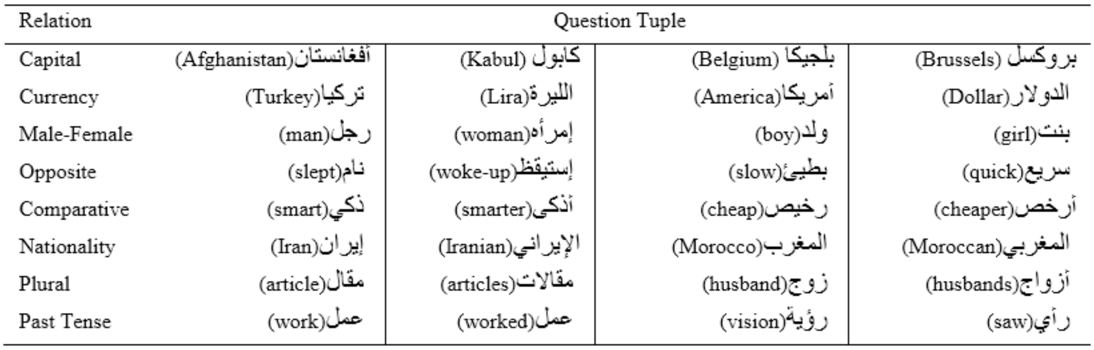
For evaluation we used an automatically POS annotated dataset by Farasa. Table 4 shows the accuracy results from our model and Aravec model using top-10 words over different analogy relations. It can be observed from the table that both the lemma-based embedding and the morphologically based embedding outperform the output from Aravec, and that the lemma-based embedding is superior to the morphological embedding. The CBOW model on wikipedia in Aravec showed the least accuracy among the three.

[2] In Arabic, definite article "el" (i.e., "the"), conjunction "wa"(i.e., "and") and "ba"(i.e., "with") are fused with the following noun as a prefix, and possessive letter "ه"(i.e., "his") is fused with the preceding noun. In this paper, we indicate that by adding a hyphen between the noun and the article, conjunction or possessive letter in the English translations.



Word	Morphological Embedding		Lemma Embedding	Aravec_wiki
(work)عمل	عمل/Verb	عمل/NOUN		
	(worked)اشتغل/V	(and-work) وعمل/N	(worked)اشتغل	(and-work) وعمل
	(and-work) وعمل /N	(work) عمل/V	(profession)مهنة	(working) يعمل
	(working) يعمل/V	(his-work) عمله/N	(activity)نشاط	(worked) اشتغل
	(served) خدم/V	(the-work) العمل/N	(experience)خبرة	(the-work) العمل
	(work) عمل/N	(working)يعمل /V	(technical)فني	(his-work) عمله
(food)أكلات	اكلات/Verb	اكلات/NOUN		
	(name of a dish) الباخمري/N	(name of a dish) شكشوك/N	(pot)كسرولة	(and-chicken) والدجاج
	(and-pears) والكمثري/N	(and-cookies) وكعك/N	(name of a dish) شكشوكة	(and-fish) والاسماك
	(with-lentils) بالعدس/N	(thick)المدمس/N	(pears)الإجاص	(and-mice) والفئران
	(name of a dish) والشخشوخة/N	(and-fruits) وفواكه/N	(name of a dish) طعمية	(insects)حشرات
	(the-pickled) المخللة/N	(and-food) واكلات/N	(with-lentils) بالعدس	(and-plants) ونباتات
(food)طعام	طعام/Verb	طعام/NOUN		
		(meal)وجبة/N	(food)غذاء	(meal)وجبه
		(the-lunch) الغداء/N	(meat)لحم	(the-lunch) الغداء
	Not in Vocabulary	(drink) شراب/N	(drink)مشروب	(the-drink) الشراب
		(lunch)غذاء/N	(taste)طعم	(bread)خبز
		(drink)الشراب/N	(milk)حليب	(the-breakfast) الافطار
(Turkey)تركيا	تركيا/Verb	تركيا/NOUN		
		(Istanbul)اسطنبول/N	(Bulgaria)بلغاريا	(Syria)سوريا
		(Ankara)انقرة/N	(Iran)ايران	(Albania)البانيا
	Not in Vocabulary	(Iran)ايران/N	(Anatol)اناضول	(Armenia)ارمينيا
		(Izmir)ازمير/N	(Azerbaijan)اذربيجان	(Bulgaria)بلغاريا
		(in-Turkey) بتركيا/N	(Armenia)ارمينيا	(Istanbul)اسطنبول

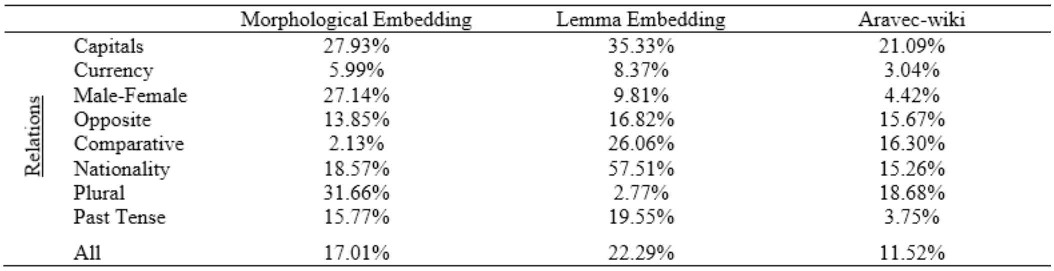
Table 2. Results from a Random Sample



Relation	Question Tuple			
Capital	(Afghanistan)أفغانستان	(Kabul) كابول	(Belgium) بلجيكا	(Brussels) بروكسل
Currency	(Turkey)تركيا	(Lira)الليرة	(America)أمريكا	(Dollar)الدولار
Male-Female	(man)رجل	(woman)إمرأه	(boy)ولد	(girl)بنت
Opposite	(slept)نام	(woke-up)إستيقظ	(slow)بطيئ	(quick)سريع
Comparative	(smart)ذكي	(smarter)أذكى	(cheap)رخيص	(cheaper)أرخص
Nationality	(Iran)إيران	(Iranian)الإيراني	(Morocco)المغرب	(Moroccan)المغربي
Plural	(article)مقال	(articles)مقالات	(husband)زوج	(husbands)أزواج
Past Tense	(work)عمل	(worked)عمل	(vision)رؤية	(saw)رأي

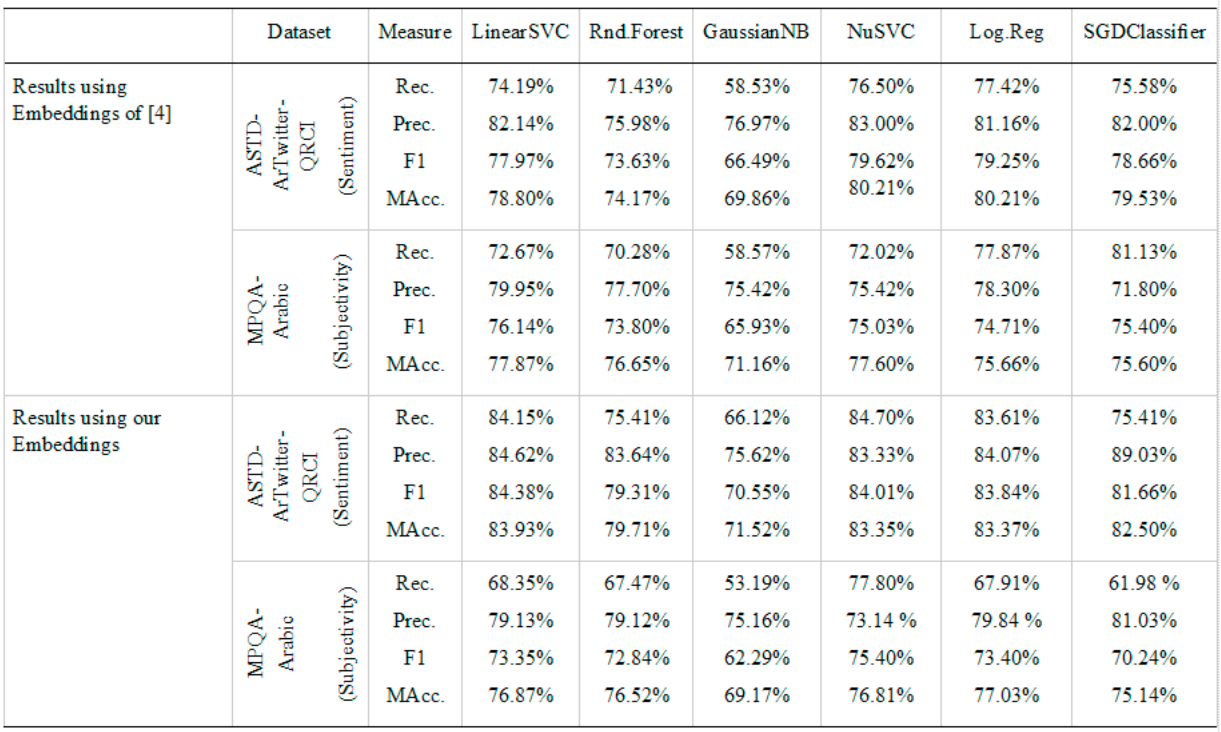
Table 3. Sample Arabic Analogy Questions

By referring to the detailed performance of lemma and morphologically based embedding on every relation separately, we can see that although their overall accuracy differs by 5.28%, their performance distribution in different relations may differ much more. For example, lemma-based embedding is better in comparative relations by 23.93% while morphologically based embedding exceeds in male-female relations by 17.33%. This strongly emphasizes our assumption that different embeddings can be used to represent different semantics at different levels.



		Morphological Embedding	Lemma Embedding	Aravec-wiki
Relations	Capitals	27.93%	35.33%	21.09%
	Currency	5.99%	8.37%	3.04%
	Male-Female	27.14%	9.81%	4.42%
	Opposite	13.85%	16.82%	15.67%
	Comparative	2.13%	26.06%	16.30%
	Nationality	18.57%	57.51%	15.26%
	Plural	31.66%	2.77%	18.68%
	Past Tense	15.77%	19.55%	3.75%
	All	17.01%	22.29%	11.52%

Table 4. Accuracy (%) Comparison Between our Models and Aravec-wiki Model.



	Dataset	Measure	LinearSVC	Rnd.Forest	GaussianNB	NuSVC	Log.Reg	SGDClassifier
Results using Embeddings of [4]	ASTD-ArTwitter-QRCI (Sentiment)	Rec.	74.19%	71.43%	58.53%	76.50%	77.42%	75.58%
		Prec.	82.14%	75.98%	76.97%	83.00%	81.16%	82.00%
		F1	77.97%	73.63%	66.49%	79.62%	79.25%	78.66%
		MAcc.	78.80%	74.17%	69.86%	80.21%	80.21%	79.53%
	MPQA-Arabic (Subjectivity)	Rec.	72.67%	70.28%	58.57%	72.02%	77.87%	81.13%
		Prec.	79.95%	77.70%	75.42%	75.42%	78.30%	71.80%
		F1	76.14%	73.80%	65.93%	75.03%	74.71%	75.40%
		MAcc.	77.87%	76.65%	71.16%	77.60%	75.66%	75.60%
Results using our Embeddings	ASTD-ArTwitter-QRCI (Sentiment)	Rec.	84.15%	75.41%	66.12%	84.70%	83.61%	75.41%
		Prec.	84.62%	83.64%	75.62%	83.33%	84.07%	89.03%
		F1	84.38%	79.31%	70.55%	84.01%	83.84%	81.66%
		MAcc.	83.93%	79.71%	71.52%	83.35%	83.37%	82.50%
	MPQA-Arabic (Subjectivity)	Rec.	68.35%	67.47%	53.19%	77.80%	67.91%	61.98 %
		Prec.	79.13%	79.12%	75.16%	73.14 %	79.84 %	81.03%
		F1	73.35%	72.84%	62.29%	75.40%	73.40%	70.24%
		MAcc.	76.87%	76.52%	69.17%	76.81%	77.03%	75.14%

Table 5. Arabic Sentiment and Subjectivity Analysis Results.

4.2. Extrinsic Evaluation

We carried out extrinsic evaluation by applying the generated embeddings to the Arabic sentiment and subjectivity analysis task in [4]. The authors built their word embeddings using a big number of local and international editions of Arabic newspapers, consumer reviews with dialectal Arabic and the complete text of the holy Quran. For subjectivity classification they used labeled news articles, while for sentiment classification they used labeled twitter datasets and book reviews. They trained six different binary classifiers to detect subjectivity and sentiment in Arabic. For a fair comparison, we used their Standard Arabic classification datasets and ignored datasets with dialectical Arabic. We fed our embeddings to their model using our unified morphological based embeddings for evaluation compatibility.
Table 4 shows a comparison between the results using the embeddings in [4] and the results using our embeddings on same datasets, we used the same naming conventions as the authors; ASTD-ArTwitter-QRCI and and MPQA-Arabic for Twitter and Arabic newspaper dataset respectively. The measures recall(Rec.), precision(Prec.), F-measure(F1), and macro-accuracy(MAcc) are used for evaluation. Additionally, we compare between their and our ROC scores using false positive against true positive rates. As shown, our model always does better on ASTD-ArTwitter-QRCI

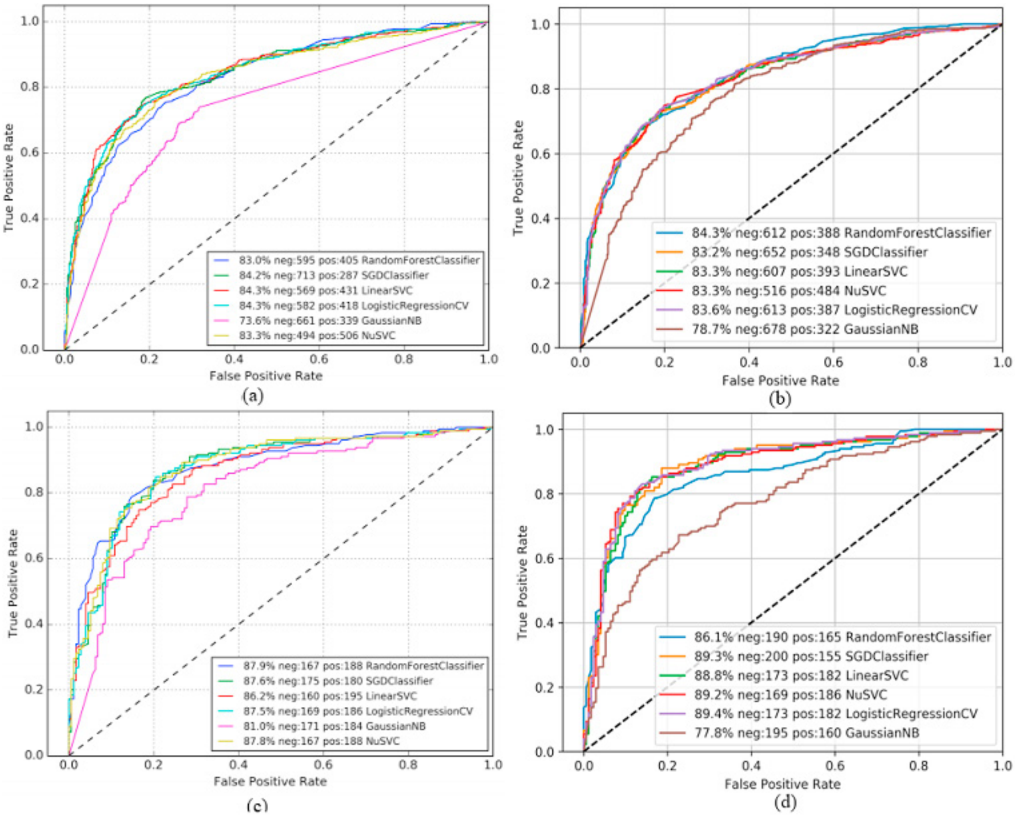


Fig. 4. (a) ROC Scores-MPQA Arabic articles[4] (b)Our Scores on MPQA Arabic articles (c) ROC Scores-ASTD-ArTwitter tweets[4] (d)Our Scores on ASTD-ArTwitter tweets.

for sentiment analysis and is comparable on MPQA-Arabic for subjectivity analysis despite the fact that we are using less data for training our word embeddings. This can also be demonstrated in Figure 5: our ROC scores on MPQA-Arabic dataset are on average very close to the Arabic sentiment results and are even better with some classifiers like, GaussianNB. Surpisingly, the ROC scores for our model on ASTD-ArTwitter-QRCI dataset gives better results on most classifiers.

5. Conclusion and Future work

This paper combines recent advances of Arabic NLP to study the interplay between morphologically based and lemma-based embeddings. We combined together recent taggers, segmenters, and benchmarks in Arabic, on the one hand, with recent advances in word embedding, on the other hand, to achieve better word representations at multiple levels. We introduced a model for embedding Arabic words using morphological fine-tuned POS tags that incorporate morphology senses . We further proposed a lemma-based embedding that represents another trade-off for word representation. Lemma-based embeddings showed very meaningful semantic representations in comparison with normal embeddings. Although they may drop morphological similarity, this may be useful for some tasks. As per our experiments, lemma embeddings may be suitable to noun words and related types like adjectives, adverbs and names, and hence more suitable for tasks like Sentiment Analysis or Named Entity Recognition. Morphological embeddings, on the other hand, may be more suitable for verbs. In general, experimental results indicate that both morphologically-based and lemma-based embeddings are quite promising for semantic representation. We achieved significant performance improvement when compared to similar models built with similar settings, and competitive results when compared to models built with much larger datasets. In general, morphologically based embedding preserves much more information in word vectors, tackling long-tail phenomena in semantic representation. Whether to

use morphologically-based or lemma-based embedding will depend on the NLP task .
In the future, we plan to train our model on much larger corpora for better coverage and higher accuracy as this was evident in the case of analogy-based evaluation. Generally, larger corpora are crucial to obtain more meaningful embeddings. We also plan to work on dialectal Arabic corpora to cover all variations of Arabic words. Additionally, we will consider out-of-vocabulary words and try to map them to the closest lemma embeddings. Moreover, we plan to perform more extrinsic evaluation to study which embeddings are more suitable for which NLP tasks. Furthermore, we plan to study embedding regeneration using morphological rules from lemma-based embeddings to represent inflected words more accurately.

References

[1] Alajmi, A., Saad, E.M., Darwish, R.R., Manning, C.D., Raghavan, P., Zou, F.Z., Wang, F.L., Deng, X., Han, S., 2012. Toward an arabic stop-words list generation.

[2] Alayba, A.M., Palade, V., England, M., Iqbal, R., 2018. Improving sentiment analysis in arabic using word representation. CoRR abs/1803.00124.

[3] AlMarwani, N., Diab, M.T., 2017. Arabic textual entailment with word embeddings, in: WANLP@EACL, Association for Computational Linguistics. pp. 185–190.

[4] Altowayan, A.A., Tao, L., 2016. Word embeddings for arabic sentiment analysis, in: 2016 IEEE International Conference on Big Data (Big Data), pp. 3820–3825. doi:10.1109/BigData.2016.7841054.

[5] Avraham, O., Goldberg, Y., 2017. The interplay of semantics and morphology in word embeddings. CoRR abs/1704.01938.

[6] Basirat, A., Tang, M., 2018. Lexical and morpho-syntactic features in word embeddings - a case study of nouns in swedish, in: Proceedings of the 10th International Conference on Agents and Artificial Intelligence - Volume 1: NLPinAI,, INSTICC. SciTePress. pp. 663–674. doi:10.5220/0006729606630674.

[7] Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., 2003. A neural probabilistic language model. JOURNAL OF MACHINE LEARNING RESEARCH 3, 1137–1155.

[8] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2016. Enriching word vectors with subword information. CoRR abs/1607.04606.

[9] Botha, J.A., Blunsom, P., 2014. Compositional morphology for word representations and language modelling. CoRR abs/1405.4273.

[10] Boudelaa, S., Pulvermüller, F., Hauk, O., Shtyrov, Y., Marslen-Wilson, W.D., 2010. Arabic morphology in the neural language system. J. Cognitive Neuroscience 22, 998–1010. URL: https://doi.org/10.1162/jocn.2009.21273, doi:10.1162/jocn.2009.21273.

[11] Collobert, R., Weston, J., 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning, in: Proceedings of the 25th International Conference on Machine Learning, ACM, New York, NY, USA. pp. 160–167. URL: http://doi.acm.org/10.1145/1390156.1390177, doi:10.1145/1390156.1390177.

[12] Cotterell, R., Schütze, H., 2015. Morphological word-embeddings, in: NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015, pp. 1287–1292. URL: http://aclweb.org/anthology/N/N15/N15-1140.pdf.

[13] Cotterell, R., Schütze, H., Eisner, J., 2016. Morphological smoothing and extrapolation of word embeddings, in: ACL (1), The Association for Computer Linguistics.

[14] Cui, Q., Gao, B., Bian, J., Qiu, S., Liu, T., 2014. Learning effective word embedding using morphological word similarity. CoRR abs/1407.1687.

[15] Dahou, A., Xiong, S., Zhou, J., Haddoud, M.H., Duan, P., 2016. Word embeddings and convolutional neural network for arabic sentiment classification, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee. pp. 2418–2427. URL: http://www.aclweb.org/anthology/C16-1228.

[16] Elrazzaz, M., Elbassuoni, S., Shaban, K.B., Helwe, C., 2017. Methodical evaluation of arabic word embeddings, in: ACL (2), Association for Computational Linguistics. pp. 454–458.

[17] Hartmann, N., Fonseca, E.R., Shulby, C., Treviso, M.V., Rodrigues, J., Aluísio, S.M., 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. CoRR abs/1708.06025.

[18] Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T., 2016. Fasttext.zip: Compressing text classification models. arXiv preprint arXiv:1612.03651 .

[19] Kuang, S., Davison, B.D., 2018. Class-specific word embedding through linear compositionality, in: 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 390–397. URL: doi.ieeecomputersociety.org/10.1109/BigComp.2018.00064, doi:10.1109/BigComp.2018.00064.

[20] Laatar, R., Aloulou, C., Belguith, L.H., 2017. Word sense disambiguation of arabic language with word embeddings as part of the creation of a historical dictionary, in: LPKM, CEUR-WS.org.

[21] Lai, S., Liu, K., Xu, L., Zhao, J., 2015. How to generate a good word embedding? CoRR abs/1507.05523. URL: http://arxiv.org/abs/1507.05523, arXiv:1507.05523.

[22] Li, Q., Shah, S., Liu, X., Nourbakhsh, A., 2017. Data sets: Word embeddings learned from tweets and general data. CoRR abs/1708.03994. URL: http://arxiv.org/abs/1708.03994, arXiv:1708.03994.

[23] Liu, Q., Ling, Z., Jiang, H., Hu, Y., 2016. Part-of-speech relevance weights for learning word embeddings. CoRR abs/1603.07695.

[24] Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013a. Efficient estimation of word representations in vector space. CoRR abs/1301.3781. URL: http://arxiv.org/abs/1301.3781, arXiv:1301.3781.

[25] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013b. Distributed representations of words and phrases and their compositionality. CoRR abs/1310.4546. URL: http://arxiv.org/abs/1310.4546, arXiv:1310.4546.

[26] Nagoudi, E.M.B., Schwab, D., 2017. Semantic similarity of arabic sentences with word embeddings, in: WANLP@EACL, Association for Computational Linguistics. pp. 18–24.

[27] Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation, in: In EMNLP.

[28] Řehůřek, R., Sojka, P., 2010. Software Framework for Topic Modelling with Large Corpora, in: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valletta, Malta. pp. 45–50. http://is.muni.cz/publication/884893/en.

[29] Shapiro, Pamelaand Duh, K., 2018. Morphological word embeddings for arabic neural machine translation in low-resource settings, in: Proceedings of the Second Workshop on Subword/Character LEvel Models, Association for Computational Linguistics. pp. 1–11. URL: http://aclweb.org/anthology/W18-1201.

[30] Soliman, A.B., Eissa, K., El-Beltagy, S.R., 2017. Aravec: A set of arabic word embedding models for use in arabic NLP, in: ACLING, Elsevier. pp. 256–265.

[31] Trask, A., Michalak, P., Liu, J., 2015. sense2vec - A fast and accurate method for word sense disambiguation in neural word embeddings. CoRR abs/1511.06388.

[32] Vulic, I., Mrksic, N., Reichart, R., Séaghdha, D.Ó., Young, S.J., Korhonen, A., 2017. Morph-fitting: Fine-tuning word vector spaces with simple language-specific rules. CoRR abs/1706.00377.

[33] Xu, Y., Liu, J., 2017. Implicitly incorporating morphological information into word embedding. CoRR abs/1701.02481.

[34] Young, T., Hazarika, D., Poria, S., Cambria, E., 2017. Recent trends in deep learning based natural language processing. CoRR abs/1708.02709. URL: http://arxiv.org/abs/1708.02709, arXiv:1708.02709.

[35] Zahran, M.A., Magooda, A., Mahgoub, A.Y., Raafat, H.M., Rashwan, M.A., Atyia, A., 2015. Word representations in vector space and their applications for arabic, in: CICLing (1), Springer. pp. 430–443.