Lang Resources & Evaluation
https://doi.org/10.1007/s10579-018-9419-x

PROJECT NOTES

The impact of corpus domain on word representation: a study on Persian word embeddings

Amir Hadifar[1] · **Saeedeh Momtazi**[1]

© Springer Nature B.V. 2018

Abstract Word embedding, has been a great success story for natural language processing in recent years. The main purpose of this approach is providing a vector representation of words based on neural network language modeling. Using a large training corpus, the model most learns from co-occurrences of words, namely Skip-gram model, and capture semantic features of words. Moreover, adding the recently introduced character embedding model to the objective function, the model can also focus on morphological features of words. In this paper, we study the impact of training corpus on the results of word embedding and show how the genre of training data affects the type of information captured by word embedding models. We perform our experiments on the Persian language. In line of our experiments, providing two well-known evaluation datasets for Persian, namely Google semantic/syntactic analogy and Wordsim353, is also part of the contribution of this paper. The experiments include computation of word embedding from various public Persian corpora with different genres and sizes while considering comprehensive lexical and semantic comparison between them. We identify words whose usages differ between these datasets resulted totally different vector representation which ends to significant impact on different domains in which the results vary up to 9% on Google analogy and up to 6% on Wordsim353. The resulted word embedding for each of the individual corpora as well as their combinations will be publicly available for any further research based on word embedding for Persian.

✉ Saeedeh Momtazi
momtazi@aut.ac.ir

[1] Computer Engineering and Information Technology Department, Amirkabir University of Technology, Hafez Avenue, Tehran, Iran

Published online: 05 July 2018

Keywords Distributional semantic models · Word embedding · Word2vec · Persian

1 Introduction

Distributional Semantic Models (DSMs), also known as word vector representation models, use vectors that keep track of context (e.g., co-occurring words) in which target terms appear in a large corpus as proxies for meaning representations, and applying geometric techniques to these vectors to measure the similarity of the corresponding words (Baroni et al. 2014). In recent years, the new generation of DSMs become so popular which called neural word embedding (word embedding trained by neural networks).

Most word embedding approaches assume each word as a single dense vector of real numbers, each number preserves some probabilistic or syntactic/semantic relationship between words. These representations are typically derived from large unlabeled corpora using co-occurrence statistics (Mikolov et al. 2013a). These vectors have been proven to be valuable features in variety of natural language processing tasks, including name entity recognition (Collobert and Weston 2008; Turian et al. 2010; Lample et al. 2016), part of speech tagging (Lin et al. 2015; dos Santos and Zadrozny 2014; Zhang et al. 2016), word sense disambiguation (Chen et al. 2014; Iacobacci et al. 2016), language modeling (Mnih and Hinton 2008; Bengio et al. 2003; Cha et al. 2017) as well as information retrieval (Zamani and Croft 2016; Zuccon et al. 2015; Kenter and de Rijke 2015; Kusner et al. 2015; Brokos et al. 2016).

There are two main approaches for learning DSMs: (1) global matrix factorization methods, such as Latent Semantic Analysis (LSA) and GLObal VEctor representation (Glove), (2) local context window methods such as, continuous Skip gram (Skip-gram) and Continuous Bag Of Words (CBOW) models (Pennington et al. 2014). Some researchers also proposed another categorization for DSMs: (1) count-based models, (2) prediction-based models (Baroni et al. 2014). In this categorization, count-based models referred to traditional DSMs such as SVD that uses optimal decomposition to solve its objective, whereas prediction-based model like Glove uses training-based approach such as stochastic gradient decent to optimize an objective.

While Pennington et al. (2014) claims that Glove outperforms other models on word analogy, word similarity, and name entity recognition but Levy et al. (2015) revealed some evidence that the main performance gains of word embedding are due to certain system design and hyper-parameter optimizations, rather than the embedding algorithms themselves. Beside this comparison, Baroni et al. (2014) compared prediction-based and count-based models and concluded that on average prediction-based models significantly outperformed count-based ones.

In this work, we consider Skip-gram as a typical word embedding and describe our results in training, hyper-parameterization and evaluation for Persian. While most of the algorithms learn word embedding according to the external context of

words and ignore internal structures, which is important limitation for morphologically rich languages, we also used proposed method by Bojanowski et al. (2017) to incorporate morphological information into word representations and show some advantage and disadvantage over conventional word embedding models. We also perform our experiments on different corpora to show how the genre of the training data affects the representation of words. In addition to these studies, a further investigation was done to compare syntactic and semantics of these models.

2 Background

Representation of words as a continuous vectors has a long history (Mikolov et al. 2013a). Recently, lots of research has been devoted to the neural network based approaches. Bengio et al. (2003), introduced a model that learn word vector representation as a part of neural network architecture for language modeling. Later, Collobert and Weston (2008) proposed a model to learn word embedding by using a multilayer neural network architecture, predicting a word based on shallow window.

In 2013, the Skip-gram and CBOW were proposed (Mikolov et al. 2013a). Both architectures are similar to neural network language modeling, but the key difference with previously mentioned approaches is lower computational complexity (for comparison, Skip-gram take less than a day to train whereas previously mentioned approaches take a week or longer on the same training data). In these two architectures, non-linear hidden layer, which was the main reason of complexity, is removed and projection layer is share for all words (Mikolov et al. 2013a).

Soon after the Skip-gram has released, same authors proposed extensions to improve quality of vector and training speed (Mikolov et al. 2013b). They replaced basic-softmax function with more sophisticated one such as Hierarchical-Softmax (HS) or Negative Sampling (NS).

In another extension by Bojanowski et al. (2017), character n-grams of words were taken into account to improve Skip-gram model. In this extension, for each word w, they include a set of character n-gram (3–6 grams) to learn word morphology too.

While most of the focus in word embedding literature is on objective function and mathematical model, other factors affect the result as well (Levy et al. 2015). Some parameters, such as window-size, vector-size, and learning rate are obviously tunable and lead to better results.

3 Related work on Persian word embedding

Despite huge research interests on word embedding in the recent years, few works have studied these models in other languages, specifically Persian.

The Polyglot project (Al-Rfou et al. 2013) trained word embedding for more than 100 languages including Persian using their corresponding Wikipedia articles. Facebook also, published 300-dimensional pretrained fastText vector for 294 languages on Wikipedia (Bojanowski et al. 2017). Gharavi et al. (2016) used word

embedding for plagiarism detection in Persian. Basirat and Joakim (2016) trained word vectors for Persian on Hamshahri using default parameter to benefit in a greedy dependency parser model. Passban et al. (2016) proposed a neural network POS tagger that input vectors generated by Word2vec and Glove which both trained on Bijankhan corpora and few vector dimensionally had been experimented. In semEval 2017 (Camacho-Collados 2017) new word similarity dataset proposed for Persian and another four languages. Participants trained their models on Wikipedia and evaluated on 500 word pairs which annotated by human annotators.

The main problem of all mentioned works in that they trained word embedding by default parameters in line with original Word2vec model; i.e., no comparison had been done between different corpora or hyper-parameters.

4 Model

In neural language modeling, language models are built based on neural network in which the probability distribution of each word w_t given its n previous words can be estimated with softmax (Bengio et al. 2003):

$$P(w_t|w_{t-1},\ldots,w_{t-n-1}) = \frac{e^{y_{w_t}}}{\sum_{i=1}^{T} e^{y_i}} \tag{1}$$

where y_t is log-probability of word w_t normalized by sum over all log-probabilities in given corpus. Therefore, the objective function tries to minimize this probability over all words T in the corpus:

$$j_\Theta = \frac{1}{T}\sum_{t=1}^{T} \log P(w_t|w_{t-1},\ldots,w_{t-n-1}) \tag{2}$$

The idea of word embedding is really tight with neural network language model. CBOW tries to predict target (center) word w_t based on sliding window of n (Fig. 1):

$$P(w_t|w_{t-n},\ldots,w_{t+n}) = \frac{e^{u^\intercal v_{w_t}}}{\sum_{i=1}^{T} e^{u^\intercal v_{w_i}}} \tag{3}$$

where u is the vector of surrounding word and v is vector of w_t as a target word. As a result, the inner product of these two vectors compute log-probability of word w_t. In fact, at training time each word has two roles, one as target word and the other as context of the other words.

In contrast to CBOW, Skip-gram tries to predict surrounding words w_{t+j} based on a target word w_t:

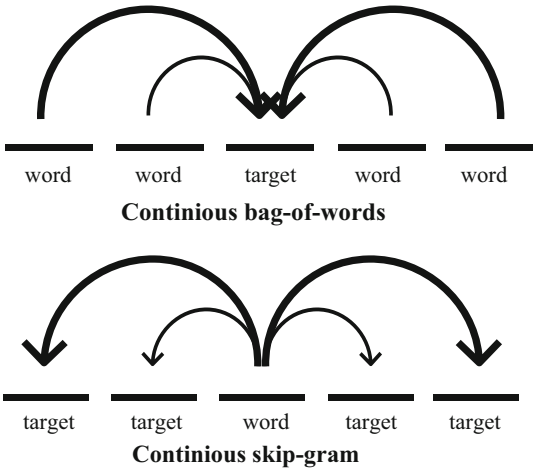

Fig. 1 Continous bag-of-words predict target based on sorrounding words, Continous skip-gram predict sorrounding targets based on current word

$$P(w_{t+j}|w_t) = \frac{e^{u^{\top} v_{w_{t+j}}}}{\sum_{i=1}^{T} e^{u^{\top} v_{w_i}}}, (-n < j < n) \tag{4}$$

but these probabilities are so expensive, as they need to sum over all words at each training step. Therefore, the basic-softmax can be replaced with more practical approaches such as HS or NS (Mikolov et al. 2013b).

HS uses a Huffman tree in which the probability distribution of words is obtained from a product of two probability for each word. This approach significantly improved calculation of denominator of softmax as only the specified path needs to be computed and updated. NS optimizes a different objective function. It is proposed to consider classification task in which a classifier is trained to discriminate between true and false examples. Instead of only using the words observed next to each other in training data as true examples, some arbitrary words sampled from the corpus as false examples.

Skip-gram with character n-grams is slightly different. In order to leverage n-grams information for each word, it denotes a set $\zeta = \{1, \ldots, G\}$ of n-grams appearing in word and computes log-probability of each word:

$$P(w_{t+j}|w_t) = \frac{e^{u^{T} z_w}}{\sum_{i=1}^{T} e^{u^{T} z_{w_i}}}, (-n < j < n) \tag{5}$$

where z_w is vector of w_t which represent as a sum over its n-grams vectors (Bojanowski et al. 2017). Therefore, log-probability of each word is calculated by product of its character n-grams embedding and context embedding.

Table 1 Persian corpora used for training

Corpora	Tokens
Persian Twitter	46,538,663
Persian Wikipedia	77,549,405
Hamshahri	148,496,668
irBlogs	864,080,384
Total	1,136,665,120

5 Data acquisition

5.1 Training corpora

We gathered our data from various Persian corpora. After a preprocessing step, which includes removing punctuations and numbers, it contains 1,136,665,120 tokens (Table 1). Our data include:

- *irBlogs* a large user generated text which crawled from 560,000+ Persian bloggers which contains nearly 5 million posts. It includes 45 topics about different subject categories which are prepared in TREC standard format. The total number of distinct words that are used in irBlogs is more than 5 million words, while in formal Persian language this number is around 680,000. The main reason for this variety in vocabulary is the use of slangs, typographical errors and abbreviations (AleAhmad et al. 2016).
- *Hamshahri* the collection are news article of Hamshahri newspaper, one of the most popular daily newspapers in Iran, from 1996 to 2002 that is written by many authors from variety of backgrounds and covers a range of different topics. It contains 16 main categories which are prepared in TREC standard format. This corpus contains more than 160,000 news articles and nearly 417,000 different words (AleAhmad et al. 2009).
- *Persian Wikipedia* the dataset consists of the entire dump of Persian Wikipedia until 01/01/2017. Articles written collaboratively by many people in various topics. It contains about 361,000 documents after pruning articles shorter than 50 words. Total number of distinct words in this corpus is around 212,000.
- *Persian Twitter* the dataset was collected through Twitter streaming API from August 2016 to February 2017. We restricted the dataset to Persian language tweets and all user-mentioned, symbols and URLs removed. Dataset contains 3,904,894 tweets and 46,538,663 tokens.

5.2 Test set

For the evaluation of Persian word embedding, two famous test set: Google analogy and Wordsim353 were developed. The Google analogy dataset were introduced by Mikolov et al. (2013a) and contains analogy questions in the form $A{:}B{::}C{:}D$; i.e., A is to B as C is to D, where the forth word (D) is missing and the developed system

Table 2 Google analogy and examples

Relation type	Example	Affected example	Unaffected
Capital common country	Baghdad:Iraq::Tehran:Iran	Islamabad (اسلام‌آباد)[a]	466/506
Capital world	Lusaka:Zambia::Tehran:Iran	Ashgabat (عشق‌آباد)	4016/4524
Currency	Canada:dollar::Iran:rial	nothing affected	866/866
City in state	Miami:Florida::Irving:Texas	Jacksonville (جسکون‌ویل)	2318/2467
Family	boy:girl::she:he	grandson (نوه‌پسر), granddaughter (نوه‌دختر)	340/506
Gram1 adjective to adverb	calm:calmly::rare:rarely	totally affected	0/992
Gram2 opposite	aware:unaware::sure:unsure	honest (راستگو), dishonest (دروغگو)	812/812
Gram3 comparative	bad:worse::big:bigger	longer (طولانی‌تر)	1332/1332
Gram4 superlative	wide:widest::bad:worst	luckiest (خوش‌شانس‌ترین)	1122/1122
Gram5 present participle	fed:feeding::fly:flying	totally affected	0/1056
Gram6 nationality adjective	India:Indian::England:English	Korean (کره‌ای)	1559/1599
Gram7 past tense	going:went::running:ran	increasing (افزایش‌دادن)	1560/1560
Gram8 plural	bird:birds::cow:cows	car (ماشین), machine (ماشین)	1332/1332
Gram9 plural verb	eat:eats::walk:walks	nothing affected	0/870

[a] Space character (␣) indicate compound words

should guess this word from the vector representation of words. To this aim, the following mathematical operation should be performed on each instance of the test data:

$$\sim v(D) = v(A) - v(B) + v(C)$$

The most similar word to $\sim v(D)$ based on cosine similarity is return as D.

The dataset contains 14 relation types as listed in Table 2. We constructed Persian equivalence through Google translation tool and correct the results manually. We omitted the relation type "adjective to adverb", "present participle", "plural verb", and some parts of other relations because they do not exist in Persian or could not be translated into a unique word. E.g. in relation type "family" we have analogy question like *she*:*he*::*boy*:*girl* which both "*she*" and "*he*" in Persian map to a single pronoun. In another example for relation type "City in State", some city names are translated to compound noun which are not capture correctly in training. In some cases we replaced phrases with more common Persian equivalence; e.g., "honest" (راستگو) and "dishonest" (دروغگو) is a good example in English to evaluate syntax, but when they are translated to Persian, syntax and structure of words are totally changed. As a result, we use another common phrase like "allowable" (مجاز) and "unallowable" (غیرمجاز) instead. Considering all these changes, the resulted Persian dataset ends to 15,763 analogies from the original 19,544 English analogies. The Persian analogies contains a total of 8006 semantic and 7757 syntactic analogies.

The Wordsim353 dataset is a collection of 353 word pairs annotated by human with their semantic relatedness (between 1 and 10) (Finkelstein et al. 2002). Using word embedding the similarity between two words can be computed with cosine similarity. Once the similarity between the words computed, we obtain the list of word pairs ordered according to these similarities which can be compared with the ordered list of words based on the similarities in gold data. Computing Spearman's correlation between these ranked lists specifies how well word embedding captures similarity. For this dataset, translation and cleaning has led to 285 pair from the

original 353 English pairs. The translation was done by three persons individually. Among the three translations provided by them, all three translators have agreement on 36.15% of pairs. For 54.38% of pairs, two translators have agreement and for 9.47% of pairs we found no agreement between translators. The gold data was provided based on the agreement between the three translations; i.e., for each pair the words which have suggested by more than one translator are selected. In case no agreement was found between the translated pairs, each of the words in the pair is selected separately based on the maximum agreement in the translations.

6 Expriment

6.1 Implementation

For creation of the Persian word embedding we used Skip-Gram with Negative Sampling (SGNS) and its character n-grams extension as training algorithms since on average they obtained best accuracy, on different test sets in distributional semantics domains (Levy et al. 2015; Bojanowski et al. 2017). We used Gensim, a python-based vector space modeling toolkit (Rehurek and Petr 2010), and fastText, a powerful C++ library for learning word representation (Bojanowski et al. 2017). Gensim is a good choice because it allows for different models to be deployed within the same framework and save them in a format compatible with the original word2vec implementation. Another advantage of Gensim is that, different corpora can be combined incrementally without a need to retrain the whole model. However, it only updates weights for existing word vectors based on existing vocabulary and it is not possible to add new vocabulary. Since we aim at having larger vocabulary by using more corpora, we combined corpora separately to get more precise evaluation. We used fastText because it is an efficient library for learning representation of words according to their morphology.[1]

6.2 Standard model

In first step of our experiments, we used default parameters of Gensim (window size = 5, vector dimension = 100, minimum count = 5, negative sample = 5 and iteration = 5) to evaluate our SGNS models on Wordsim353 (Table 3) and Google analogy (Table 4). The main purpose of this experiment is to study the impact of training corpora on the achieved results. To this aim, the model is trained on each of the gathered corpora separately and combined interchangeably. The number of resulted word vectors can be seen in Table 5.

We tried to keep the experiments as unsupervised as possible, because Persian is a low-resource language that suffers from efficient preprocessing toolkits. Therefore no normalization or stemming is used.

As can be seen in the results of Wordsim353 test set, Wikipedia and Hamshahri achieved the best results compared to other individual corpora due to proper size

[1] We will make our scripts and models available upon publication.

Table 3 SGNG Spearman's correlation on different corpora. The highest value in each combination is marked in bold

Corpora	Spearman's correlation
Twitter	0.4547
Wikipedia	**0.5398**
Hamshahri	0.5293
irBlogs	0.5165
Twitter + Wikipedia	0.5635
Twitter + Hamshahri	0.5609
Wikipedia + Hamshahri	**0.5673**
Wikipedia + irBlogs	0.5206
Hamshahri + irBlogs	0.5360
Twitter + Wikipedia + Hamshahri	**0.5691**
Wikipedia + Hamshahri + irBlogs	0.5370
All corpora	0.5494

and less noisy content. Although Twitter could not achieve good results individually, combining this data with Hamshahri or Wikipedia improved the results due to capturing different concepts. This indicates that capturing different concepts and genres is an important factor in achieving better results.

The most obvious observation, as expected—is obtaining better accuracy when we combined datasets even when we added noisy ones like irBlogs or Twitter. In a total comparison, we observed that combination of Wikipedia, Hamshahri and Twitter lead to best results.

To compare the effect of capturing larger data or different genres we performed another experiment on Twitter and Wikipedia. In this experiment, we selected half of the Twitter corpus and trained the word2vec model with this data. The Spearman's correlation was 0.4235 (compared to 0.4547 on the whole Twitter data). We then selected the same amount of data from Wikipedia and combined it with our dataset provided from half of Twitter. The resulted dataset has equal size to the original Twitter corpus, but the Spearman's correlation on this mixed dataset was 0.5263, significantly higher that the Twitter data alone. The results of this experiment indicate that the having more variation in content and genre of data is very important to achieve better performance.

Table 4 presents the ratio of correctly discovered analogies in each dataset. Wikipedia discovered more analogy than others (7945 entries), because of its high quality and its relevance to analogy test set. Relation types such as "city in state" or "currency" not captured as well as English due to lack of context about these topics in Persian corpora; e.g., in a large dataset such as irBlogs there are only 6 entries discovered out of 2318 for "city in state" relation types.

Best semantic result achieved by Wikipedia with 46.2% accuracy between individual corpora, but it's not performed very well in syntactic types. On the other side, irBlogs captured syntactic feature better than others, with 51.9% accuracy, but it performed poorly on semantic analogies. The same results also observed when combining each of these two corpora with other ones. The best overall accuracy achieved when we combined both of them.

Table 4 SGNS evaluation on Google analogy

Analogy type	Twitter	Wikipedia	Hamshahri	irBlogs	Twitter + Wikipedia	Twitter + Hamshahr
Capital common country	16.2% (44/272)	77.9% (360/462)	42.2% (195/462)	30.4% (73/240)	75.3% (348/462)	45.3% (172/380)
Capital world	14.6% (58/398)	58.5% (1256/2147)	41.3% (637/1544)	16.2% (65/400)	56.5% (922/1632)	40.5% (521/1286)
Currency	0.8% (2/240)	8.3% (35/420)	11.9% (50/420)	2.5% (6/240)	11.7% (49/420)	6.4% (22/342)
City in state	13.3% (2/15)	22.5% (224/994)	6.9% (11/160)	0% (0/6)	23% (126/548)	4.8% (3/63)
Family	53.9% (97/180)	40.8% (97/238)	54.3% (113/208)	65% (117/180)	56.2% (117/208)	59.6% (124/208)
Gram2 opposite	5.2% (11/210)	10.4% (19/182)	18.6% (86/462)	22.5% (54/240)	13.8% (33/240)	20.8% (71/342)
Gram3 comparative	29.7% (124/418)	35.3% (178/504)	56.7% (397/700)	71.9% (582/810)	55.4% (279/504)	64.7% (356/550)
Gram4 superlative	28.1% (76/270)	27% (73/270)	29.4% (111/378)	39.1% (119/304)	41.9% (113/270)	44.4% (151/340)
Gram6 nationality adjective	27.2% (343/1260)	58.9% (919/1560)	55.5% (781/1406)	45.5% (541/1190)	64.4% (905/1406)	57% (801/1406)
Gram7 past tense	36.8% (412/1120)	26.5% (246/928)	22% (247/1121)	57.2% (721/1260)	43.9% (492/1120)	31.1% (369/1188)
Gram8 plural	11% (20/182)	15.8% (38/240)	27.1% (103/380)	43.1% (199/462)	27.6% (75/272)	33.7% (128/380)
Accuracy on semantic	18.1% (203/1105)	46.2% (1972/4261)	36% (1006/2794)	24.4% (261/1066)	47.7% (1562/3270)	36.9% (842/2279)

Table 4 continued

Analogy type	Twitter	Wikipedia	Hamshahri	irBlogs	Twitter + Wikipedia	Twitter + Hamshahr
Accuracy on syntactic	28.4% (986/3460)	39.9% (1473/3684)	38.7% (1725/4447)	51.9% (2216/4266)	49.7% (1897/3812)	44.6% (1876/4206)
Overall accuracy	26% (1189/4565)	43.4% (3445/7945)	37.7% (2731/7241)	46.5% (2477/5332)	48.8% (3459/7082)	41.9% (2718/6485)

Analogy type	Wikipedia + Hamshahri	Wikipedia + irBlogs	Hamshahri + irBlogs	Twitter + Wikipedia + Hamshahri	Wikipedia + Hamshahri + irBlogs	All corpora
Capital common country	76.6% (354/462)	47.1% (144/306)	34.9% (95/272)	64.1% (296/462)	73.8% (177/240)	57% (195/342)
Capital world	56.7% (1086/1917)	27.2% (178/654)	24.3% (148/610)	53.2% (919/1727)	42.2% (169/400)	38.7% (337/871)
Currency	14.5% (61/420)	6.6% (18/272)	7.2% (22/306)	12.1% (51/420)	6.2% (15/240)	9.4% (32/342)
City in state	18.3% (100/547)	1.5% (1/66)	0% (0/15)	15.1% (64/423)	50% (3/6)	4.4% (2/45)
Family	50.4% (120/238)	59.1% (123/208)	64.4% (116/180)	58% (138/238)	53.3% (96/180)	63.9% (115/180)
Gram2 opposite	19.5% (74/380)	21% (57/272)	23.2% (63/272)	22.5% (77/342)	18.3% (44/240)	20.3% (62/306)
Gram3 comparative	60.7% (334/550)	70.6% (494/700)	75.5% (569/754)	65.1% (358/550)	52.6% (426/810)	75.1% (526/700)
Gram4 superlative	35.9% (122/340)	46.1% (140/304)	47% (143/304)	42.8% (130/304)	29.9% (91/304)	45.1% (137/304)
Gram6 nationality adjective	69% (1022/1482)	57.8% (728/1260)	51.3% (610/1190)	70% (1038/1482)	64% (762/1190)	63.4% (845/1332)

Table 4 continued

Analogy type	Wikipedia + Hamshahri	Wikipedia + irBlogs	Hamshahri + irBlogs	Twitter + Wikipedia + Hamshahri	Wikipedia + Hamshahri + irBlogs	All corpora
Gram7 past tense	27.6% (309/1120)	56.9% (717/1260)	53.4% (673/1260)	35.8% (425/1188)	41.7% (525/1260)	54.4% (685/1260)
Gram8 plural	35.8% (136/380)	47% (217/462)	46.3% (214/462)	33.4% (127/380)	26.2% (121/462)	41.9% (176/420)
Accuracy on semantic	48% (1721/3584)	30.8% (464/1506)	27.5% (381/1383)	44.8% (1468/3270)	43.1% (460/1066)	38.2% (681/1780)
Accuracy on syntactic	46.9% (1997/4252)	55.2% (2353/4258)	53.5% (2272/4242)	50.7% (2155/4246)	46.1% (1969/4266)	56.2% (2431/4322)
Overall accuracy	47.4% (3718/7836)	48.9% (2817/5764)	47.2% (2653/5625)	48.2% (3623/7516)	45.6% (2429/5332)	51% (3112/6102)

Table 5 Number of word vectors and tokens

Corpora	Tokens	Word vectors
Twitter	46,538,663	129,185[a]
Wikipedia	77,549,405	153,308
Hamshahri	148,496,668	127,305
irBlogs	864,080,384	625,907
Twitter + Wikipedia	124,088,068	239,570
Twitter + Hamshahri	195,035,331	211,538
Wikipedia + Hamshahri	226,046,073	221,708
Wikipedia + irBlogs	941,629,789	685,562
Hamshahri + irBlogs	1,012,577,052	662,750
Twitter + Wikipedia + Hamshahri	272,584,736	298,713
Wikipedia + Hamshahri + irBlogs	1,090,126,457	719,612
All corpora	1,136,665,120	758,585

[a] Number of word vectors for each corpus is the same for all experiments

As can be seen in the tabulated results, the behavior of these corpora varies for different ontologies. For example, higher portion of family relations are captured by irBlogs, while the number of capital relations captured by Wikipedia or the number of currency relations captured by Hamshahri is significantly higher than other corpora.

As stated above, Wikipedia captures semantic feature better than irBlogs but in "family" relation type, irBlogs beaten Wikipedia by a large margin. The reason is "family" relation type is less common in Wikipedia articles, while in everyday speaking we use "brother", "father" and other "family" relation type much more.

In another case, analogy type "gram7 past tense" captured in irBlogs and Twitter better than Wikipedia and Hamshahri, as it is more common to use past tense in everyday speech rather than Wikipedia which mainly describe facts and usually written in present perfect. Overall comparison of results implies that, user generated data captured syntactic features better than semantics ones.

The other observation is about "currency" relation type. Newspapers such as Hamshahri, usually report latest currency news or exchange rate between different currencies. As a result this kind of relation is discovered in Hamshahri better than other corpora.

This observation shows the difference between the content of these corpora and indicates that the application domain must be taken into the consideration when selecting training data.

6.3 Semantic differences

In the next step of our experiments, we study the semantic differences between these corpora in more detail. To this aim, we focused on ambiguous words in Persian; i.e.,

Table 6 Semantic difference between Wikipedia and irBlogs

Word	Meaning	Most similar words in Wikipedia	Most similar words in irBlogs
Astin (آستین)	a) The capital of the U.S. state of Texas b) Part of garment that covers a person's arm	Powers (پاورز), Dallas (دالاس), Amarillo (آمارلو), Houston (هوستون)	sleeve (استین), sleeves (استینهای), a sleeve (استینی), his/her sleeve (آستینهایش)
Kelk (کلک)	a) An imaginary pen that is used in Persian poetry b) Deceive or outwit (someone)	opal (نگین), spite(لج), dandelion (قاصدک), Chista (چیستا)	and trick (وکلک), bag (جوال), knack (حقه), deception (نیرنگ)
Ootoo (اتو)	a) Small hand held appliance, is used to remove creases b) An American theatre and film director	Otto (اتو), Preminger (پرمنیجر), Ludwig (لودویگ), Rodolph (رودولف)	cloth-iron (اطو), hair-dryer (سشوار), a cloth-iron (اتوی), ironing (اتوکشی)
Kabk (کبک)	a) City in Canada b) A brown bird with a round body and short tail	Riviere (ریویر), Missisquoi (میسیسکوا), Laurentides (لورانتید), Chaudiere (شودیر)	Pheasant (قرقاول), a Perdicinae (کبکی), Ammoperdix griseogularis (تیهو), Sandgrouse (باقرقره)

Table 7 SGNG with character n-gram evaluation on Google analogy

Analogy type	Twitter	Wikipedia	Hamshahri	irBlogs	Twitter + Wikipedia	Twitter + Hamshahri
Capital common country	30.5% (83/272)	83.5% (386/462)	71.6% (331/462)	40.8% (98/240)	75.1% (347/462)	64.7% (246/380)
Capital world	20.4% (81/398)	66.6% (1429/2147)	64.9% (1002/1544)	26.8% (107/400)	57.5% (941/1632)	56.3% (724/1286)
Currency	0.4% (1/240)	9.5% (40/420)	13.6% (57/420)	4.2% (10/240)	10% (42/420)	10.5% (36/342)
City in state	6.7% (1/15)	22.5% (235/1045)	7.5% (12/160)	0% (0/6)	17.7% (97/548)	4.8% (3/63)
Family	49.4% (89/180)	29.4% (70/238)	52.4% (109/208)	66.1% (119/180)	50% (104/208)	64.9% (135/208)
Gram2 opposite	39% (82/210)	13.2% (24/182)	29.2% (135/462)	34.2% (82/240)	30.8% (74/240)	31% (106/342)
Gram3 comparative	61% (255/418)	41.9% (211/504)	65% (455/700)	69.4% (562/810)	65.7% (331/504)	72.4% (398/550)
Gram4 superlative	63.7% (172/270)	34.4% (93/270)	52.4% (198/378)	47% (143/304)	57.8% (156/270)	69.7% (237/340)
Gram6 nationality adjective	83% (1046/1260)	86.6% (1351/1560)	86.3% (1214/1406)	66.9% (796/1190)	84.1% (1182/1406)	82.1% (1154/1406)
Gram7 past tense	65.1% (729/1120)	44.9% (417/928)	40.7% (456/1121)	56.4% (711/1260)	63% (706/1120)	52.9% (629/1188)
Gram8 plural	47.3% (86/182)	25% (60/240)	42.6% (162/380)	45.2% (209/462)	54% (147/272)	46.8% (178/380)
Accuracy on semantic	23% (255/1105)	50% (2160/4312)	54% (1511/2794)	31.3% (334/1066)	46.8% (1531/3270)	50.1% (1144/2279)

Table 7 continued

Analogy type	Twitter	Wikipedia	Hamshahri	irBlogs	Twitter + Wikipedia	Twitter + Hamshahri
Accuracy on syntactic	68.4% (2370/3460)	58.5% (2156/3684)	58.9% (2620/4447)	58.6% (2503/4266)	68.1% (2596/3812)	64.2% (2702/4206)
Overall accuracy	57.5% (2625/4565)	54% (4316/7996)	57.1% (4131/7241)	53.2% (2837/5332)	58.3% (4127/7082)	59.3% (3846/6485)

Analogy type	Wikipedia + Hamshahri	Wikipedia + irBlogs	Hamshahri + irBlogs	Twitter + Wikipedia + Hamshahri	Wikipedia + Hamshahri + irBlogs	All corpora
Capital common country	76.8% (355/462)	58.8% (180/306)	42.6% (116/272)	73.8% (341/462)	58.5% (200/342)	55.3% (189/342)
Capital world	65.5% (1255/1917)	45% (294/654)	39% (238/610)	63.9% (1103/1727)	50.7% (442/871)	49.3% (429/871)
Currency	12.6% (53/420)	4.4% (12/272)	6.2% (19/306)	13.6% (57/240)	5.3% (18/342)	7% (24/342)
City in state	22.3% (122/547)	13.6% (9/66)	0% (0/15)	22.5% (95/423)	11.1% (5/45)	11.1% (5/45)
Family	50.8% (121/238)	70.2% (146/208)	65.6% (118/180)	55.9% (133/238)	68.3% (142/208)	70% (126/180)
Gram2 opposite	28.2% (107/380)	32% (87/272)	37.5% (102/272)	31% (106/342)	31% (95/306)	34.6% (106/306)
Gram3 comparative	66% (363/550)	70% (490/700)	72.7% (548/754)	72.5% (399/550)	72.9% (510/700)	74.4% (521/700)
Gram4 superlative	51.2% (174/340)	49.3% (150/304)	47.4% (144/304)	57.6% (175/304)	51.3% (156/304)	55.3% (168/304)
Gram6 nationality adjective	86.5% (1282/1482)	74.1% (934/1260)	69.9% (832/1190)	84.8% (1257/1482)	78.4% (1044/1332)	77.1% (1027/1332)

Table 7 continued

Analogy type	Wikipedia + Hamshahri	Wikipedia + irBlogs	Hamshahri + irBlogs	Twitter + Wikipedia + Hamshahri	Wikipedia + Hamshahri + irBlogs	All corpora
Gram7 past tense	42% (470/1120)	60.6% (764/1260)	56.7% (714/1260)	55.4% (658/1188)	56.9% (717/1260)	58.3% (734/1260)
Gram8 plural	45% (171/380)	44.8% (207/462)	46.1% (213/462)	43.4% (165/380)	45% (189/420)	49.8% (209/420)
Accuracy on semantic	53.1% (1906/3584)	42.5% (641/1506)	35.5% (491/1383)	55.9% (1729/3090)	44.6% (807/1808)	43.4% (773/1780)
Accuracy on syntactic	60.3% (2567/4252)	61.8% (2632/4258)	60.1% (2553/4242)	65% (2760/4246)	62.7% (2711/4322)	63.9% (2765/4322)
Overall accuracy	57.1% (4473/7836)	56.8% (3273/5764)	54.1% (3044/5625)	59.7% (4489/7516)	57.4% (3518/6130)	58% (3538/6102)

Table 8 SGNG with character n-grams Spearman's correlation on different corpora. The highest value in each combination is marked in bold

Corpora	Spearman's correlation
Twitter	0.5327
Wikipedia	0.5911
Hamshahri	**0.5968**
irBlogs	0.5848
Twitter + Wikipedia	0.5842
Twitter + Hamshahri	0.5880
Wikipedia + Hamshahri	**0.6245**
Wikipedia + irBlogs	0.5922
Hamshahri + irBlogs	0.5882
Twitter + Wikipedia + Hamshahri	**0.6210**
Wikipedia + Hamshahri + irBlogs	0.5937
All corpora	0.6045

words with same syntax but totally different meanings; e.g. the word "آستین" (Astin) which have two different senses: (1) Austin, the state capital of Texas, (2) sleeve, the part of a garment that covers arms.

As mentioned, Wikipedia and Hamshahri has written in formal language and mainly includes facts or news, in contrast to irBlogs and Twitter which are mostly written informally. This difference between the nature of training corpora affected the semantic meaning of words captured by their vector representation. To represent this issue, in Table 6, we listed some of the ambiguous Persian words as well as the list of most similar words captured by Wikipedia and irBlogs; e.g., closest words to "آستین" (Astin) in Wikipedia is based on the first sense of this word (Austin), while in irBlogs, the similar words capture the second sense of this word (sleeve). In most of the cases, ambiguous words in Wikipedia referred to named entities such as person or location while in irBlogs referred to their general meanings (see other example in Table 6).

6.4 Character embedding model

In the third step of our experiments, we trained our models (SGNG with character n-grams) using fastText, with the same parameters as before. The first important observation was that fastText do significantly better in capturing syntactic relation (Table 7), since it incorporate character n-grams. As a comparison with basic SGNG model, the accuracy of the relation type "nationality adjective" in all corpora improved up to 31%.

Improvement of the results using character n-grams has been achieved on Wordsim353 test set as well (Table 8). Comparing different corpora for this experiment ends to similar observations as basic SGNG model.

In another experiment we compared the most similar words extracted with respect to word embedding and char embedding. Using char embedding, give lower priority to co-occurrence feature and higher to n-gram; e.g., most similar words to "فیروزکوه" (Firozkoh, name of location), share sub-word "فیروز" (feerooz) or "کوه"

Table 9 Lexical comparison between basic SGNG and its extension

Word	Most similar words based on SGNG with character n-gram	Most similar words based on SGNG
Firozkoh (فیروزکوه)	Firozkohian (فیروزکوهی), from mountain (زکوه), from mountain (از کوه), Firozkola (فیروزکلا), green mountain (سبزکوه)	Chalus (چالوس),[a] Pakdasht (پاکدشت), Damavand (دماوند), Roudehen (رودهن), Damghan (دامغان)
library (کتابخانه)	and library (وکتابخانه), in library (درکتابخانه), librarý (کتابخانهٔ), the libraries (کتابخانههای), library (کتابخاه)	A library (کتابخانهٔ), librarý (کتابخانهٔ), archive (آرشیو), museum (موزه), file (بایگانی)
Alchemy (کیمیا)[b]	an alchemy (کیمیای), el alchemy (الکیمیا), chemical (کیمیاوی), alchemies (کیمیاهای)	Sanaz (ساناز),[c] Malihe (ملیحه), Mahtab (مهتاب), Ghazale (غزاله), Azita (آزیتا)
hunter (شکارچی)	the hunts (شکارهای), hunts (شکارها), hunters (شکارچیانی), the hunts (شکارهایی), hunters (شکارچیان)	bait (طعمه), hog (گراز), hunters (شکارچیان), tame (اهلی), animals (حیوانات)

[a] They are name of cities/locations close to Firozkoh

[b] Also a name for girls in Persian

[c] They are name of girls

Table 10 Window size effect on different corpora

Corpora	2	5	8	10	15	20	25
Twitter	0.4378[a]	0.4643	0.4724	0.4875	0.5122	0.5233	0.5227
	28.1%	31.5%	32.6%	32.8%	34%	35.3%	35%
Wikipedia	0.5059	0.5455	0.5921	0.5841	0.6024	0.6047	0.6022
	38.5%	50.8%	52.4%	53.1%	54.1%	53.3%	53.1%
Hamshahri	0.4958	0.5246	0.5560	0.5697	0.5810	0.5815	0.5913
	42.9%	48.4%	51.7%	51.8%	53.6%	53.3%	54.1%
irBlogs	0.4895	0.5180	0.5180	0.5180	0.5180	0.5180	0.5180
	51.7%	53.5%	52.9%	52.6%	52.2%	51.9%	50.4%

[a] Spearman's correlation and Google analogy respectively

Table 11 Vectors dimensional effect on different corpora

Corpora	100	200	300	400	500	1000	2000
Twitter	0.4547	0.4646	0.4628	0.4643	0.4681	0.4595	0.4662
	26%	30.8%	32.1%	30.1%	31.2%	30.3%	29.6%
Wikipedia	0.5398	0.5496	0.5516	0.5498	0.5556	0.5595	0.5511
	43%	48.8%	51.2%	49.3%	49.8%	49.7%	49.6%
Hamshahri	0.5293	0.5293	0.5235	0.5288	0.5332	0.5261	0.5344
	37.7%	46.9%	48.7%	49.7%	49.4%	49.7%	49.6%
irBlogs	0.5165	0.5250	0.5180	0.5213	0.5228	0.5120	0.5008
	46.5%	51.2%	53.5%	53%	54.3%	53.9%	53%

Table 12 Negative samples effect on different corpora

Corpora	2	4	8	12	16	20	30
Twitter	0.4661	0.4602	0.4718	0.4618	0.4628	0.4635	0.4748
	28%	30.2%	32.4%	32%	32.7%	31.6%	30.2%
Wikipedia	0.5611	0.5576	0.5515	0.5581	0.5474	0.5585	0.5514
	45.9%	49.4%	51.5%	51.2%	51.2%	50.2%	50%
Hamshahri	0.5307	0.5228	0.5317	0.5367	0.5359	0.5297	0.5365
	46.9%	48.3%	47.7%	47.9%	47.1%	47.4%	45.6%
irBlogs	0.5253	0.5166	0.5232	0.5284	0.5335	0.5282	0.5335
	54.5%	53.5%	52.6%	51.8%	51.8%	50.7%	49.6%

(mountain) while using basic SGNG, most similar words are name of other places and not have any sub-word in common (see other example in Table 9).

6.5 Hyper parameters

In last step of our experiments, we evaluated the effect of vector dimension, sliding window and number of negative samples on different corpora. These three parameters have huge impact on quality of word vectors.

As stated in original word2vec paper, larger window size of training examples lead to higher accuracy, at the expense of the training time (Mikolov et al. 2013b), therefore (as shown in Table 10) with increase of window size accuracy improved. In all cases we have significant improvement from window size of 2–5 which implies that minimum efficient window size start around 5. It is hard to say which window size is perfect for each dataset but it seems to be around 5–15. In this observation, we found with growth of window size syntactic accuracy decreased while semantic accuracy (almost) not changed. This is probably because in larger window we lose useful syntactic structure.

Another important parameter in word embedding is vectors dimensions. As it can be expected both using more data and higher dimensional word vectors will improve the accuracy. The strongest result for Wikipedia, appearing around 300, while in irBlogs because of its larger size it appearing around 500 (Table 11). The result obtained for the vectors dimensional seem to be line with those obtained for English (Mikolov et al. 2013a).

The least but not the last hyper-parameter is number of negative samples. As reported in original paper (Mikolov et al. 2013b) for small dataset, negative samples between 5 and 20 are useful and for larger dataset, this number can be around 2–5. For small dataset like Twitter this number is around 16 and for irBlogs 2–3 seemed to be good (Table 12).

7 Conclusion

In this paper, we have provided vector representation for Persian words based on word2vec model using SGNG and the combination of SGNG with character n-grams. To this aim, we used various corpora to study the impact of training data of the results achieved when capturing semantic or syntactic features of words. To perform evaluation, two popular test sets, namely Wordsim353 and Google analogy, were translated and adopted for Persian.

We compared semantic and syntactic differences between our corpora and observed each of these corpora is good in capturing particular relation types. Moreover, we showed how the differences in the training data leads to capturing different senses of ambiguous words.

We believe that our work suggests room for further improvements in Persian word embedding. In future we plan to improve Google analogy and replace missing analogies with related ones.

Acknowledgements Twitter dataset provided by Ali Shariat Bahadori from university of Tehran. Any usage and statement made herein are solely the responsibility of the authors.

References

Al-Rfou, R., Perozzi, B., & Skiena, S. (2013). Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of CoNLL* (pp. 183–192).

AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M., & Oroumchian, F. (2009). Hamshahri: A standard persian text collection. *Knowledge Based Systems*, *2*, 382–387.

AleAhmad, A., Zahedi, M. S., Rahgozar, M., & Moshiri, B. (2016). IrBlogs: A standard collection for studying Persian bloggers. *Computers in Human Behavior*, *57*, 195–207.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL* (pp. 238–247).

Basirat, A., & Joakim, N. (2016). Greedy universal dependency parsing with right singular word vectors. In *Proceedings of SLTC*.

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*, 1137–1155.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Brokos, G., Malakasiotis, P., & Androutsopoulos, I. (2016). Using centroids of word embeddings and word mover's distance for biomedical document retrieval in question answering. In *proceedings of BioNLP* (pp. 114–118).

Camacho-Collados, J., et al. (2017). Semeval-2017 Task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of SemEval2017*.

Cha, M., Gwon, Y., & Kung, H. T. (2017). Language modeling by clustering with word embeddings for text readability assessment. In *Proceedings of CIKM* (pp. 2003–2006).

Chen, X., Liu, Z., & Sun, M. (2014). A unified model for word sense representation and disambiguation. In *Proceedings of EMNLP* (pp. 1025–1035).

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing. In *Proceedings of ICML* (pp. 160–167).

dos Santos, C. N., & Zadrozny, B. (2014). Learning character-level representations for part-of-speech tagging. In *Proceedings of ICML* (pp. 1818–1826).

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., et al. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, *20*, 116–131.

Gharavi, E., Bijari, K., Zahirnia, K., & Veisi, H. (2016). A deep learning approach to persian plagiarism detection. In *Proceedings of FIRE* (pp. 154–159).

Iacobacci, I., Pilehvar, M. T., & Navigli, R. (2016). Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of ACL* (pp. 897–907).

Kenter, T., & de Rijke, M. (2015). Short text similarity with word embeddings. In *Proceedings of CIKM* (pp. 1411–1420).

Kusner, M. J., Sun, Y., Kolkin, N. I., & Weinberger K. Q., et al. (2015). From word embeddings to document distances. In *Proceedings of ICML* (pp. 957–966).

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT* (pp. 260–270).

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211–225.

Lin, C.-C., Ammar, W., Dyer, C., & Levin, L. (2015). Unsupervised POS induction with word embeddings (pp. 1311–1316).

Mikolov, T., Chen, K., Corrado, G., Dean, J.(2013a). Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS* (pp. 1–9).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013b). Efficient estimation of word representations in vector space (pp. 1–12). arXiv:1301.3781v3.

Mnih, A., Hinton, G. E. (2008). A scalable hierarchical distributed language model. In *Proceedings of NIPS* (pp. 1–8).

Passban, P., Qun, L., & Way, A. (2016). Boosting neural POS tagger for Farsi using morphological information. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *16*, 1–15.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of EMNLP* (pp. 1532–1543).

Rehurek, R., & Petr, S. (2010). Software framework for topic modelling with large corpora. In *Proceedings of LREC* (pp. 45–50).

Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. In *Proceedings of ACL* (pp. 384–394).

Zamani, H., & Croft, W. B. (2016). Embedding-based query language models. In *Proceedings of ICTIR* (pp. 147–156).

Zhang, Y., Gaddy, D., Barzilay, R., & Jaakkola, T. S. (2016). Ten pairs to tag-multilingual pos tagging via coarse mapping between embeddings. In *proceedings of NAACL-HLT* (pp. 1307–1317).

Zuccon, G., Koopman, B., Bruza, P., & Azzopardi, L. (2015). Integrating and evaluating neural word embeddings in information retrieval. In *Proceedings of ADCS* (pp. 1–8).