

---

# SAT BASED ANALOGY EVALUATION FRAMEWORK FOR PERSIAN WORD EMBEDDINGS

---

A PREPRINT

**Seyyed Ehsan Mahmoudi**  
Shahid Beheshti University  
se\_mahmoudi@sbu.ac.ir

Mehrnoush Shamsfard  
Shahid Beheshti University  
m-shams@sbu.ac.ir

July 1, 2021

## ABSTRACT

In recent years there has been a special interest in word embeddings as a new approach to convert words to vectors. It has been a focal point to understand how much of the semantics of the the words has been transferred into embedding vectors. This is important as the embedding is going to be used as the basis for downstream NLP applications and it will be costly to evaluate the application end-to-end in order to identify quality of the used embedding model. Generally the word embeddings are evaluated through a number of tests, including analogy test. In this paper we propose a test framework for Persian embedding models. Persian is a low resource language and there is no rich semantic benchmark to evaluate word embedding models for this language. In this paper we introduce an evaluation framework including a hand crafted Persian SAT based analogy dataset, a colloquial test set (specific to Persian) and a benchmark to study the impact of various parameters on the semantic evaluation task.

## 1 Introduction

In recent years Word Embedding has been used extensively in many NLP applications. Specially with vast use of deep learning methods, having an embeddings for words, has been a de-facto standard. This highlights the importance of embedding methods and in the same time urges for a proper framework for evaluating the embedding models. One wholistic approach is to evaluate the embedding models in an end-to-end eco system as extrinsic evaluation method, which will use different embedding models in the specific application and chooses the best one. But this approach is going to be costly in terms of the required computation power. In addition to all hyper-parameters that need to be tuned in later stages it will be challenging to pin point the end-to-end performance for embedding model.

Alternatively we can have some intrinsic test frameworks to evaluate the quality of embedding model. Traditionally Analogy, Similarity and Categorization tests has been incorporated to perform such intrinsic model evaluations. Analogy tests try to compare the relation of two pairs of words such as *King( $v_1$ ) to Queen( $v_2$ ) is like Man( $v_3$ ) to ?* . This type of test is solved by finding the nearest neighbor of  $v_3 + v_2 - v_1$ . Word Similarity tests, on the other hand, try to assign a similarity score to pair of words, which normally is the cosine similarity to the words and compare it to human assigned score. Another category of intrinsic tests are Word Categorization tests which try to run a clustering on the words and evaluate the quality of the formed clusters. The test datasets that have been used so far, are mainly inspired by the initial analogy test set that has been used in original Word2Vec and SGNS papers [Mikolov et al., 2013]. This test set has been challenged that it does not cover many semantic aspects of the language. Plus the fact that almost 50 percent of the dataset is dedicated to some narrow information such as capital-country and country-currency tests. Although solving such analogies shows the power of the embedding model, but in the same time it does not reflect the quality of the embedding model on more semantic oriented fashion. In this sense, there are arguments that the Google analogy dataset is not as challenging as the SAT (Scholastic Aptitude Test) dataset [Church, 2017]. Inspection shows that the SAT analogies are all semantic (not syntactic) and involve relatively complex relations. It has been shown that although the answer of analogy is in the top answers but if we inspect other words in top n nearest neighbor, they are not relevant

Table 1: SAT example question

<b>Stem</b>	<i>mason:stone</i>
a	<i>teacher:chalk</i>
b	<i>carpenter:wood</i>
c	<i>soldier:gun</i>
d	<i>photograph:camera</i>
e	<i>book:word</i>
<b>Solution</b>	<i>carpenter:wood</i>

choices. For example in the *King to Queen is Like Man to ?*, if we inspect top 10 nearest neighbors we expect the majority to be female nouns but this not true [Church, 2017]. SAT questions are solved pretty much like the analogy test but reflect much deeper semantic relations. SAT multiple-choice questions come as a stem which is pair of words and 5 other pairs that we should select the pair that has the same semantic relation as the stem. Example of such tests are shown in Table 1.

The above problem also exists in Persian language model evaluations where try to simulate the same analogy and word similarity tests on Persian words [Zahedi et al., 2018, Hadifar and Momtazi, 2018]. The analogy test set that is used, has 16 categories of relations where 11 of them are syntactic or morphological relations (See Table 3) [Zahedi et al., 2018]. It does not explore the other aspects and complexity of semantic relations that are crucial in NLP tasks. In addition to that, Persian language has some special characteristics that are not relevant in other languages such as Colloquialism or colloquial language which has important footprint in Persian language. In this paper we are going to propose a dataset to evaluate colloquial aspect of the models. We are going to explore the impact of training corpus on performance of the trained models to show how the various aspects of the Persian language will be captured based on the nature of the corpus.

## 2 Analogy Task

Analogy holds between two word pairs:  $a \rightarrow a^* \equiv b \rightarrow b^*$  ( $a$  is to  $a^*$  as  $b$  is to  $b^*$ ) For example, Tokyo is to Japan as Paris is to France. With the pair-based methods, given  $a \rightarrow a^* \equiv b \rightarrow ?$ , the task is to find  $b^*$  from rest of the corpus. An alternative method is to use the set-based methods. With set-based methods, the task is to find  $b^*$  given a set of other pairs (excluding  $b \rightarrow b^*$ ) that hold the same relation as  $b \rightarrow b^*$ . In NLP analogies [Zweig et al., 2013] are interpreted broadly as basically any "similarities between pairs of words", not just semantic. See [Church, 2017] analysis of Word2Vec, which argues that the Google analogy dataset is not as challenging as the SAT dataset. Inspection shows that the SAT analogies are all semantic (not syntactic) and involve relatively complex relations. In [Jurgens et al., 2012a] we can see a taxonomy of the relations that are used in GRE exam and are quite extensive in terms of covering various relationship types. Table 2 shows a long list of various semantic relations that can be used in analogy task. When the analogy task in Google dataset has accuracy between 80-97 percent in various categories. For SAT analogies this value drops to less than 10 percent. Which shows a huge gap on complexity and completeness of SAT question set.

It has been shown that solving analogies in SAT is far harder for embedding models [Shnayder et al., 2004] [Turney, 2008]. Same paper also explores the other top N answers of the analogy questions to show that other top 10 answers are not as impressive of the correct answer that we are looking for. This is important as in the real applications we don't know the correct answer and if we look at top N and get a good result, what is the guarantee for the first result to be the best one. In English language we know that solving SAT questions is considered to be hard problem that human weighted voting has around 80 percent of accuracy [Lofi, 2013].

In this paper we have constructed a SAT test benchmark for Persian language. The categories and taxonomy of the relations are kept but the test data set is rebuilt from scratch to reflect true and genuine relations in the words. As the majority of the semantic relations are quite deep, simple translation will definitely fail to produce acceptable outcome.

### 2.1 Colloquial Relations

Another aspect of building a good benchmark for Persian language is to consider the specific language aspects and reflect them in our test data. One of the specific aspects that we have to pay special attention is the difference in Colloquial and written form of Persian language. The difference between colloquial and written Persian is much deeper than the difference between colloquial and written English [Ghomeshi, 2018, Hamzeh and Chen, 2018]. In recent years by wide-spread of social media in Persian speaking society there has been a significant shift in Colloquial and written form of Persian [Shohani and Hosseini, 2018]. Almost no NLP application can work without supporting the Colloquial

Table 2: Semantic relations taxonomy (Note that Persian examples are Left to Right)

Class	Sub-Class	English Example
Class Inclusion	Taxonomic	emotion:rage
Class Inclusion	Functional	weapon:knife
Class Inclusion	Singular Collective	clothing:shirt
Class Inclusion	Class Individual	mountain:Everest
Part-Whole	Object:Component	car:engine
Part-Whole	Collection:Member	forest:tree
Part-Whole	Mass:Portion	water:drop
Part-Whole	Event:Feature	wedding:bride
Part-Whole	Activity:Stage	shopping:buying
Part-Whole	Item:Topological Part	mountain:foot
Part-Whole	Object:Stuff	salt:sodium
Part-Whole	Creature:Possession	robin:nest
Part-Whole	Item:Distinctive Nonpart	horse:wings
Part-Whole	Item:Ex-part/Ex-possession	apostate:belief
Similar	Synonymity	buy:purchase
Similar	Dimensional Similarity	stream:river
Similar	Dimensional Excessive	concerned:obsessed
Similar	Dimensional Naughty –	copy:plagiarize
Similar	Conversion	apprentice:master
Similar	Attribute Similarity	painting:movie
Similar	Coordinates	son:daughter
Contrast	Contradictory	masculinity:femininity
Contrast	Contrary	thin:fat
Contrast	Reverse	buy:sell
Contrast	Directional	front:back
Contrast	Defective	fallacy:logic
Attribute	Item:Attribute (noun:adjective)	soldier:wounded
Attribute	Object Attribute:Condition (adjective:adjective)	brittle:broken
Attribute	Agent Attribute:State (adjective:noun)	contentious:quarrels
Attribute	Object:Typical Action (noun:verb)	glass:break
Attribute	Agent/Object Attribute:Typical Action (adjective:verb)	mutable:change
Non-Attribute	Item:Non-Attribute (noun:adjective)	harmony:discordant
Non-Attribute	Object Attribute:Noncondition (adjective:adjective)	exemplary:criticized
Non-Attribute	Object:Nonstate (noun:noun)	famine:plentitude
Non-Attribute	Attribute:Nonstate (adjective:noun)	immortal:death
Non-Attribute	Object:Atypical Action (noun:verb)	recluse:socialize
Case Relations	Agent:Object	tailor:suit
Case Relations	Agent:Recipient	doctor:patient
Case Relations	Agent:Object - Raw Material	baker:flour
Case Relations	Action:Object	tie:knot
Case Relations	Action:Recipient	teach:student
Case Relations	Object:Recipient	speech:audience
Case Relations	Object:Instrument	violin:bow
Cause-Purpose	Cause:Effect	joke:laughter
Cause-Purpose	Cause:Compensatory Action	hunger:eat
Cause-Purpose	Instrument:Intended Action	gun:shoot
Cause-Purpose	Cause-Purpose Enabling Agent:Object	car:gas
Cause-Purpose	Action/Activity:Goal	Education:Learning
Space-Time	Item:Location	arsenal:weapon
Space-Time	Location:Process/Product	bakery:bread
Space-Time	Location:Action/Activity	school:learn
Space-Time	Time:Action/Activity	summer:harvest
Space-Time	Attachment	belt:waist
Reference	Sign:Significant	siren:danger
Reference	Expression	smile:friendliness
Reference	Plan	agenda:meeting
Reference	Knowledge	psychology:mind
Reference	Concealment	code:meaning

Table 3: Classic Relations in Analogy Dataset

Relationship Type	Number
Family Relationship	342
Currency	1260
Country-Capital	5402
Province-Capital	7832
Adjective-Adverb	1332
Noun-Adverb	1056
Antonym	1260
Comparative	1260
Superlative	1260
Nationality	1406
Singular-Plural	2550
1st Person	1260
3rd Person	1332
Infinitive-Past	1260
Infinitive-Present	1260

form in online applications. This leads us to the fact that for the evaluation framework we need to consider this aspect of the language which is missing from main stream NLP datasets on English language.

### 3 Previous Work

Unfortunately there has not been much done on exploration of embedding models for Persian. In [Zahedi et al., 2018, Hadifar and Montazi, 2018] the performance models are assessed in Analogy, Word Similarity and Categorization tasks. As in this paper we are focusing on analogy based evaluation methods, we will have a closer look at analogy datasets that has been used. The test sets that has been used for analogy, assess the model in categories that are mentioned in 3. As it is clear very little attention has been paid to finer semantic details of the language also the Colloquial aspect is completely missing from the dataset.

## 4 The proposed framework

In this paper we are trying to find a more semantic oriented framework for evaluating the embedding models in Persian language. We first describe the process of the test data creation, which is basically a two step process. We first build a categorical related words dictionary and for the second step we extract a pool of questions based on the categorical relational dictionary. Once the dataset is ready we will benchmark the result on a wide range of models with various hyper parameters to assess and analyse the framework.

### 4.1 Categorical Relational Dictionary

The first step in our SAT framework is to create the questions. As for English language there is a long history of exams such as GRE(Graduate Record Examinations), building such datasets is much easier [Turney et al., 2003]. In Persian on the other hand we chose not to translate the English datasets. As the relations have deep semantic aspects, translation becomes misleading. This is due to the fact that deep semantic relations normally rely on specific sense of the words and for majority of the words, building a one to one relation between the words in two languages is not possible or practical. Furthermore disambiguation of word senses need them to be in context, while analogy tests use separate words and so using WSD methods before translation is not possible. On the other hand due to some lexical, conceptual or cultural gaps, we may not find suitable Persian correspondence for each English pair in the dataset. For this reason, we build a new hand crafted categorical word dictionary from scratch for Persian Language. We based our work on relational categories proposed for SemEval2017 [Jurgens et al., 2012a]. The relational categories are quite cross lingual and are properly relevant to Persian language. We create word pairs as dictionary entries belonging to all of the relational categories. We created word pairs by human experts and we required at least 5 pairs per each category. We build 390 word pairs that are organized in 67 relational categories mentioned in Table 2. Those are the same categories that are specified in [Jurgens et al., 2012b]. Because in our tokenization methods we are tokenizing based on spaces, our models are incapable of working with compound verbs. For this limitation, for compound verbs we replaced them with their infinitive which is just one token. This keeps the semantic context of the word and in the same time overcomes the

Table 4: Training Corpora Details

Source	Sentences	Token Count
Wikipedia	5M	55M
Persian Blogs	400+M	5.4 Billion
Persian Twitter	70M	930M

tokenization limitation. On the same note for some of the words we did not stick to POS of words that are in the table and focused on the semantic relation that is represented by the categories. Again this was due to semantic aspects of the Persian languages that we take into consideration.

#### 4.1.1 Building the Test Pool

The next step is to build the test pool based on the categorical word pairs. We use a randomized algorithm to create a comprehensive question pool. Each 5-choice question has a stem which is a word pair and 5 answer options which are word pairs as well. Each question has just one correct answer and the word pair in correct answer option is in the same relational category of the stem word pair and other pairs are selected from other categories. Using randomized algorithm we generate questions. As mentioned above the actual number of word pairs in our dataset is 390 but the combinations of word pairs against other categories will make a much larger space for questions to be generated.

We prepared a question set for automatic evaluation of the models containing 5000 questions. From this dataset we use 1000 randomly drawn subset for human cross validation. The human cross validation dataset is to assess the human ability to solve such test. The reason to use a subset of question was to meet the attention span of our audience. We reduced the test size for each person to 20 questions. For each participant in our experiment we sequentially provided the questions to them and collected the answers.

For each question we have a pair of words as stem,  $a \rightarrow a^*$  and a number (5) of options  $o_i \rightarrow o_i^*$ . We simply calculate vectors  $a^* - a$  and options vectors  $o_i^* - o_i$  and calculate the cosine similarity of options vectors with stem vector. The option with smallest distance is considered to be the answer of the question. In this task the base-line performance is random choice that has 20 percent chance to be correct. So the random baseline is 20 percent. The algorithm is quite straight forward and no hyperparameter exists.

## 4.2 Colloquial Analogy

Also to explore the behavior of embedding models in colloquial aspects of language, specific dataset has been prepared that only contains such pairs of words. As there is no formal rules to convert words to their colloquial counterpart and also the meaning is the same on two words this relation is treated specially. The analogy data that has been prepared has 2332 analogy questions and it is as the same form as traditional analogy datasets. We intentionally created a separate dataset for this part as it does not fit into the SAT questions category.

## 5 Experimental Setup

### 5.1 Training Corpora

We want to explore impact of the base corpus on performance of the model on two tasks (SAT Analogy and Colloquial Analogy) that are introduced above. We use corpora from Persian Wikipedia, Persian Blog Corpus (hmBlog) [Motahari and Shamsfard, 2020] and Persian Twitter Corpus. Each of the mentioned sources is representing various aspects of Persian Language. In Table 4 you can find more details on the corpora that is used for training. All of the above corpora are normalized and cleaned up. The characters not in Persian alphabet are removed (except the numeric characters). The multi-form characters that are similar in written form but different unicode characters, has been transformed to their standard Persian Unicode character. This normally happens once the original texts are typed in Arabic keyboard layouts. For this reason the training corpora are normalized.

### 5.2 Model Training

There are various methods to train the embedding model. we will be examining FastText (CBOW and Skipgram with Negative Sampling SGNS) methods [Bojanowski et al., 2016]. FastText used as a representative of many methods that are used in the literature. For all of them we are exploring dimensions (50, 100, 300, 400) and window size (3, 5,

Table 5: Classic Analogy Result

Model	Corpus	D	W	SAT	Analogy	Colloquial	
CBOW	Wikipedia	100	3	0.385	0.290	0.068	
			5	0.374	0.277	0.063	
			10	0.348	0.270	0.060	
		300	3	0.394	0.258	0.093	
			5	0.390	0.262	0.078	
			10	0.374	0.251	0.076	
		400	3	0.391	0.233	0.087	
			5	0.387	0.231	0.083	
			10	0.356	0.232	0.073	
	Blogs	100	3	0.421	0.301	0.431	
			5	0.407	0.296	0.386	
			10	0.396	0.309	0.332	
		300	3	0.439	0.356	0.450	
			5	0.424	0.350	0.401	
			10	0.413	0.362	0.342	
		400	3	0.436	0.350	0.450	
			5	0.428	0.355	0.399	
			10	0.417	0.376	0.345	
	Twitter	100	3	0.360	0.183	0.396	
			5	0.344	0.184	0.364	
			10	0.342	0.183	0.343	
		300	3	0.382	0.180	0.375	
			5	0.373	0.186	0.363	
			10	0.366	0.178	0.331	
		400	3	0.387	0.176	0.351	
			5	0.379	0.172	0.336	
			10	0.366	0.171	0.315	
	SGNS	Wikipedia	100	3	0.379	0.397	0.083
				5	0.373	0.395	0.064
				10	0.363	0.377	0.040
			300	3	0.395	0.370	0.099
				5	0.388	0.394	0.103
				10	0.368	0.418	0.064
			400	3	0.397	0.315	0.080
				5	0.388	0.348	0.077
				10	0.384	0.405	0.075
Blogs		100	3	0.434	0.406	0.548	
			5	0.428	0.412	0.506	
			10	0.410	0.411	0.451	
		300	3	<b>0.473</b>	0.457	<b>0.602</b>	
			5	0.453	0.463	0.582	
			10	0.434	0.479	0.549	
		400	3	0.472	0.466	0.578	
			5	0.441	0.467	0.567	
			10	0.437	<b>0.481</b>	0.533	
Twitter		100	3	0.361	0.250	0.515	
			5	0.350	0.255	0.474	
			10	0.342	0.251	0.458	
		300	3	0.404	0.218	0.414	
			5	0.385	0.212	0.437	
			10	0.373	0.231	0.444	
		400	3	0.394	0.193	0.333	
			5	0.394	0.190	0.341	
			10	0.380	0.202	0.370	

10). Considering 3 corpora that will be used for training, it ended up in 72 trained models. On each model 3 tasks are running, normal analogy test[Zahedi et al., 2018], SAT-like analogy test and colloquial analogy test.

## 6 Experiment Results

Table 5 outlines the results of all three tasks on the trained models. There is a significant difference in the results of the various models. This shows the significant impact of both corpus and the hyper parameters that we have to choose on the final outcome.

### 6.1 Analogy Task

Here also the impact of the corpus is clear and simply can be stated that the larger corpus results in the better result. The best result for Analogy (**0.48**) improves the previous benchmarks by **3%** [Zahedi et al., 2018] which is clearly the impact of the larger corpus used in this article. The result of the less diverse corpora (Wikipedia and Twitter) are significantly lower compared to hmBlog corpus we used. Same as what has been reported before SGNS outperforms the CBOW models [Zahedi et al., 2018, Hadifar and Momtazi, 2018].

In analogy task the variation is much less compared to other tasks. This signifies that the test itself is not giving us a proper indication of the quality of the model in subsequent tasks. As the variation is high and values are quite close to each other.

Table 6: Tasks correlation Analysis

SAT - Analogy	0.37
Colloquial - Analogy	0.20
SAT - Colloquial	0.59

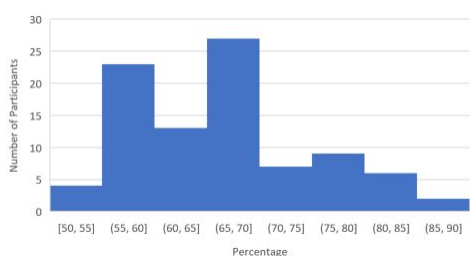


Figure 1: Human Voting Result Histogram

## 6.2 Colloquial

For Colloquial analogy task, it is clear that formal text resources for corpus such as Wikipedia are performing really poor. The highest accuracy achieved on Wikipedia corpus was **0.10** compared to other models that achieved up to **0.60** for Colloquial analogy task. This shows the importance choosing proper corpus to capture specific aspects of the language.

## 6.3 SAT Test

In the SAT test it is clear that the result is heavily dependent on the quality of the corpus and parameters. Interesting point is that majority of the output are having almost the same accuracy but for specific models the outcome is significantly better. This shows the fact that this test can be a good discriminatory measure for various embedding models. Even on a rich corpus such as hmBlog, the diversity of the results is quite notable. The best result belongs to SGNS model trained on hmBlogs corpus with dimension 300 and window size 3. Although for both Analogy and SAT the best corpus is the hmBlog corpus but the best result comes out of a different hyper parameters. It is interesting that the best result for SAT and Colloquial Analogy are based on the same model. Although the colloquial analogy test has the same format as the classic analogy test, but as it is more inclined to the semantic aspect of the language, the colloquial analogy results show more correlation with SAT compared to classic analogy. Table 6 shows that SAT test has the highest mutual correlation of the results with the other methods.

## 7 Human SAT Analogy Test

In order to find the difficulty level of the task of analogy for Persian language we developed a test website <sup>1</sup> to conduct the test by human and evaluate the general difficulty of the task. For this experiment, as mentioned before, 1000 random questions formed the main question bank. We provided 20 questions to each participant. The data was collected and stored in database. Once participants start the test, they are given the next batch of the questions. We did not perform any randomization here to make sure that all questions will be answered at least once.

In total 94 participants took part in the test. The average result of all participants was 68 percent with median of 70 percent. In Figure 1 you can find the histogram of the result of the human answers. This quite resides in the same range of English SAT test historical result which is 57 percent [Shohani and Hosseini, 2018]. This indicates the complexity and ambiguity of the task even for human. In the same time outlines the richness of the embedding models that have achieved quite human comparable outcome.

## 8 Conclusion

The results of various tasks specific for Persian language demonstrate that special attention needs to be made for characteristics of the language being processed. Also we see that the coverage of corpus plays important role on

<sup>1</sup><https://sbu-nlp-sat.herokuapp.com/>

performance of the embedding models in the tasks. As shown the SAT test result had better mutual correlation with the other two test datasets. This highlights that it can be used as good measure of the model quality. Introduction of the SAT based analogy task, helps us assess the depth of the semantic richness of the embedding model before jumping into broader end-to-end application. Looking at the results of human performance versus the embedding model performance signifies the maturity and richness of the embedding models. Also the dataset created in this research can act as a benchmark on Persian language, whilst unfortunately such datasets are rare. For next steps we are working to extend the dataset beyond the current hand crafted list of words to more automated way of extracting deep semantic relations. One of the good candidates that we are working on is the FarsNet [Shamsfard et al., 2010]. This helps us to develop much broader dataset with larger words in it.

## References

- [Bojanowski et al., 2016] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.
- [Church, 2017] Church, K. W. (2017). Word2vec. *Natural Language Engineering*, 23(1):155–162.
- [Ghomeshi, 2018] Ghomeshi, J. (2018). 12 the associative plural and related constructions in persian. *Trends in Iranian and Persian Linguistics*, 313:233.
- [Hadifar and Momtazi, 2018] Hadifar, A. and Momtazi, S. (2018). The impact of corpus domain on word representation: a study on persian word embeddings. *language resources and evaluation*, 52(4):997–1019.
- [Hamzeh and Chen, 2018] Hamzeh, M. and Chen, J. (2018). A contrastive analysis of persian and english vowels and consonants. *Lege Artis*, 3(2):105–131.
- [Jurgens et al., 2012a] Jurgens, D., Holyoak, K. J., Mohammad, S. M., and Turney, P. D. (2012a). Semeval-2012 task 2: Measuring degrees of relational similarity. *joint conference on lexical and computational semantics*, 1:356–364.
- [Jurgens et al., 2012b] Jurgens, D. A., Turney, P. D., Mohammad, S. M., and Holyoak, K. J. (2012b). 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*, page 356–364, USA. Association for Computational Linguistics.
- [Lofi, 2013] Lofi, C. (2013). Just ask a human? - controlling quality in relational similarity and analogy processing using the crowd. *btw workshops*, pages 197–210.
- [Mikolov et al., 2013] Mikolov, T., Sutskever, I., Corrado, G. S., Dean, J., and Chen, K. (2013). Distributed representations of words and phrases and their compositionality. *arXiv: Computation and Language*.
- [Motahari and Shamsfard, 2020] Motahari, H. and Shamsfard, M. (2020). Extracting metaphorical adjective phrases based on corpus and word embedding models. In *2nd National Conference on Applied Research in Computational Linguistics*.
- [Shamsfard et al., 2010] Shamsfard, M., Hesabi, A., Fadaei, H., Mansoory, N., Famian, A., Bagherbeigi, S., Fekri, E., Monshizadeh, M., and Assi, M. (2010). Semi automatic development of Farsnet: the persian wordnet. In *Proceedings of 5th Global WordNet Conference*, volume 29.
- [Shnayder et al., 2004] Shnayder, V., Littman, M. L., Bigham, J. P., and Turney, P. D. (2004). Combining independent modules in lexical multiple-choice problems. *recent advances in natural language processing*, pages 101–110.
- [Shohani and Hosseini, 2018] Shohani, A. and Hosseini, S. (2018). The impact of cyberspace on contemporary persian language and literature. *Persian Language and Literature*, (237):75–101.
- [Turney, 2008] Turney, P. D. (2008). The latent relation mapping engine: algorithm and experiments. *Journal of Artificial Intelligence Research*, 33(1):615–655.
- [Turney et al., 2003] Turney, P. D., Littman, M. L., Bigham, J., and Shnayder, V. (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. *CoRR*, cs.CL/0309035.
- [Zahedi et al., 2018] Zahedi, M. S., Bokaei, M. H., Shoeleh, F., Yadollahi, M. M., Doostmohammadi, E., and Farhoodi, M. (2018). Persian word embedding evaluation benchmarks. In *Electrical Engineering (ICEE), Iranian Conference on*, pages 1583–1588. IEEE.
- [Zweig et al., 2013] Zweig, G., tau Yih, W., and Mikolov, T. (2013). Linguistic regularities in continuous space word representations. *north american chapter of the association for computational linguistics*, pages 746–751.