International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Domain-Specific View of Semantic

Yousif Haroon[1], Mohsen Rshwan[2]

[1]Computer Science and Information Technology College, Sudan University of Science and Technology, Khartoum, Sudan

[2]Electronics and Communications Department, Cairo University, Egypt

Abstract: ***This paper investigates how universal pre-trained embedding can be adapted tohandle the diversity of semantic in specific domains.And present a novel adapting method that relies on the view of semantic constraints to adapt an off the shelfwordembedding to close fit the view of the semanticsof the target domain. The view of semantic constraints are extracted from quite available resources, a text corpus and a domain-specific dictionary. The view of semantic extraction eliminates the need for rare special-purpose semantic resources and the additional effort for locally trained embedding. The methodis implemented as a lightweight final tuning process.The results show that our method outperformed the state of the art embedding adapting methods in the task of Community Question Answering (CQA) for the Arabic medical domain.***

Keywords: Word Embedding, Embedding Adaptation, Query Expansion, Transfer Learning, NLP

1. Introduction

In word embedding new trend studies, a high quality pre-trained embedding is oriented to handle various specific domains efficiently[1-3]which represent a hot research topic in modern natural language processing. The point is that a direct adoption of a high quality universal pre-trained embedding like Word2Vec[4, 5], GloVe[6] or FastText[7] in specific domain tasks leads to poor performance[8, 9] and on the other hand, the use of locally trained embedding is only limited to its target domain [3].

Methods follow this line of research is either inject semantic lexicon similarity information into pre-trained embeddings such as Counter-Fitting[10] and Specialization[11] or combine a domain-specific embedding with pre-trained embeddings such as Sarma et al. [3] and Yin et al. [12]. However, special purpose semantic lexicons are rarely available for specific domains in the Arabic Language [22] and training domain-specific embedding locally requires an appropriate selection of training data and adequate parameter tuning[8].

We propose an adapting method that adapts pre-trained word embedding to close fit the view of the semantic of any specific domain based on the view of semantic of the target domain. The view of semantic specifies how similarity and association semantic relations are associated with the target domain[13]. Similarity relation individually gets much attention than association relationin embedding adapting methods such as Paragram[14], Counter-Fitting [10], Specialization [11] and Deep Extrofitting[15]. However, in semantics handling similarity and association jointly is preferable, because they are highly associated[13].

One way to realize the problem of word embedding is that, words can have serval views of semantic according to its domain if it is news, sport, or medical…etc. and only word embedding can capture a single view of semantics. In other words, the view of semantics in universal pre-trained embedding is general,which will not fit the view of the semantics of a specific domain, and the view of semantic in locally trained embedding is special only for specific domain and it will not match any special view for other specific domains.

Our method is based on the view of semantics constraints extracted from domain corpus to adapt a single pre-trained embedding to handle various specific domains according to the view of the target domain.The local view of semantic constraints is traditionally extracted from semantic lexicon for a special purpose[10, 11, 14, 16]. In contrast, our proposed method extracts these constraints from a text corpus and a dictionary of key terms of the target domain. The extracted view of semantics constraints represents the jointrelationship between the similarity and association relations in the given domain[13]. The method goal is to adjust the semantic representation of pre-trained embedding based on the extracted view constraints.

In order to achieve the goal, ourmethod adjusts the view of semantic inpre-trainedembeddingby tightening the word vectors pairs for each word pairs appear in the list of an extracted view of semantic constraints. The method employs the attract part of the specialization algorithm [11]and implemented as a post-processing step. Our method is also related to transfer learning[2], which assumed a new trend in embedding research work.

Experiments are conducted onSemEval-2016 Task-3 Arabic subtask-D[17], which is a re-ranking task for a Community Question Answering (CQA) for the Arabic medical domain. The task highlights the semantic techniques in CQA problems for Arabic, and it also provides the research community with abenchmark dataset for training, development, and testing plus evaluation tools [17].

For thetask mentioned above, we proposed a method based on semantic resources and embedding based query expansion techniques to expand the user question and to re-rank the relative 30 Question-Answers (QA) pairs for each question in the development dataset. Expansion sets are extracted two types of word embeddings a universal pre-trained embedding for Arabic FastText-ar[7], and a locally trained embedding for SemEval2016 task3-D training. Both embeddings are adapted using special view constraints extracted from Arabic medical domain dictionary webtib.com and the training corpus for SemEval-2016 Task-

3 Arabic subtask-D [17].

The results show that an adapted pre-trained word embedding using local view of semantic constrains out performed the locally trained embedding, and the locally trained embedding has a significant improvement in the performance after tuned using special view constrains of its specific domain.

Although our goal focuses on pre-trained embedding, the locally trained embedding also gets benefited when tuned using the local view of semantic constraints of the target domain, Figure1a,b.shows Tensorflow[18] embedding visualization for locally trained embedding before and after tuning using the view of semantic constraintsin the Arabic medical domain.

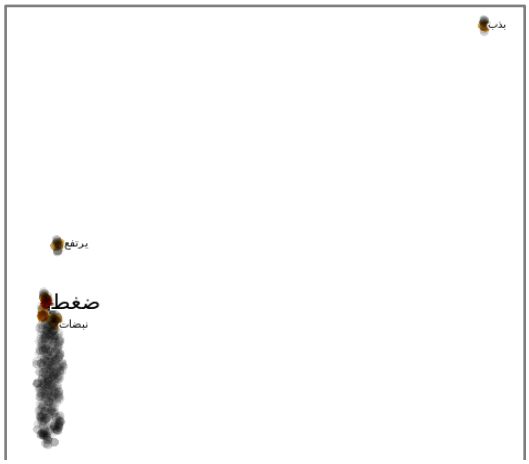

Figure 1a: Before Tuned

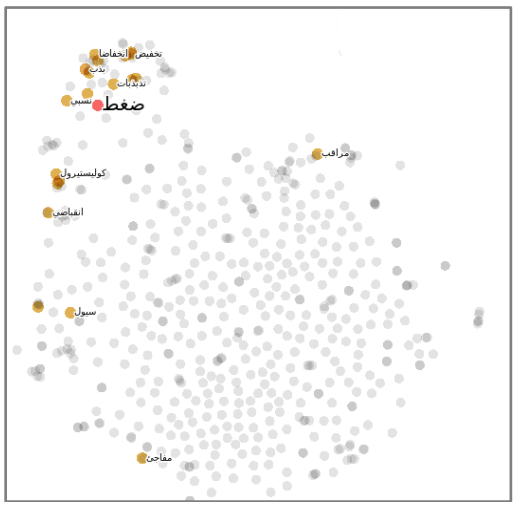

Figure 1b: After Tuned

2. Related Work

Set The purpose of adjusting the semantic representation in word embedding can fall into one of the two following categories:

2.1 Word Embedding Enrichment

Methods in this category rely on an extracted similarity or dissimilarity relation constraints from an external semantic resource like WordNet[19], Paraphrase Database (PPDB) [20], and BabelNet[21] to adjust the semantic representation of word embedding by pulling the embedding of similarity word pairs close to each other and bush dissimilarity word pairs far from each other. The aim is to enrich embedding by injecting semantic resource information. This approach attracts several embedding research works including Paragram[14], Counter-Fitting [10], Specialization [11] and Deep Extrofitting[15]. However, this approach focuses only on the similarity relation information between words and ignores the association relation [11].

2.2 Word Embedding Adaptation

This approach adapts the universal pre-trained embedding to satisfy the semantic properties of the target-specific domain. The aim is to make universal pre-trained embedding efficiently used in various specific domains [2].

The common method in this approach combines universal pre-trained embedding with locally trained embedding using techniques such as CCA Algorithm in Sarma et al. [3] and SVD in Yin et al.[12] methods. However, this approach requires the existence of locally trained embedding which requires enough training dataset and tuning effort to training model[9].

Asr et al. [8] proposed another method that averaging the first-word embedding with the second-word context embedding for each associated word pairs in the list of association constraints extracted from an external semantic resource. However, the main limitation of this model is that context embedding is ignored in most embedding techniques [8].

Our method falls into the second category, it extracts the view of semantic constraints from a widely available resource like a domain-specific dictionary and text corpus, then uses these constraints to adapt an off the shelf embedding using an efficient and lightweight process.

The second category is also related to a new trend in embedding research work the transfer learning [2], in which an abstract pre-trained neural embedding is adapted using an output layer which is trained using the target domain corpus as final fine-tune step[2].

3. The view of Semantics

In a distributional vector space model, the semantic similarity and opposite relations are generally viewed as a set of close together words or a set of far apart words respectively [7]. This generic view can have various forms based on the domain to which the training data is associated to [3], so word embedding explicitly capture similarity relation and implicitly the association relation, furthermore in recent embedding research works the similarity gets much attention than association relation [8], however, similarity and association relations are highly associated [13]. Inspired from this we introduce the view of semantic to handle both relations jointly. The view of semantic focuses on the joint between a set of similar words i.e. words that appear in the same context [7]. And their association to that context i.e. for any of the similar words, the context word comes to the mind next [23] and vice

versa. See Figure 2. In which the two head arrows represent similarity relation which symmetric and the association relation with one head arrows represent association relation which asymmetric relation [23].

To specify the view of semantics for specific domain, our method rely on a domain-specific dictionary to determine the key terms of the target domain, and then specify the words that appear previous or next to each key term, the aim is to specify the locality of the meaning of the current term which is dynamically changed according to the domain in use.

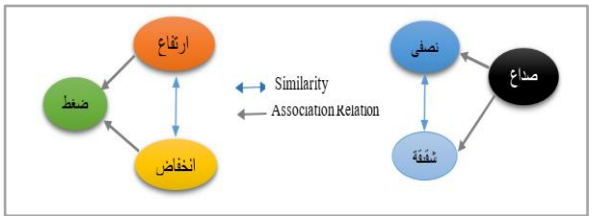

Figure 2: The View of Semantics for Some Words in Arabic Medical Domain

4. The Special view of Semantic Extraction

If The special view of semantic constraints is traditionally extracted from a specific domain semantic lexicon. However, this lexical based approach generally is limited in coverage [15] and also such semantic resource is rarely available for non-English Language like Arabic [22].

Due to this situation we propose a method to extract the special view constraints based on a domain-specific dictionary plus a text corpus of the target domain, the method is widely applicable because both resources are quit available for most domains, for example, webteb.com provides a dictionary for diseases and Drugs in the Arabic medical domain.

Our method first uses a domain-specific dictionary of the target domain to specify the key terms in the domain and then extract the most frequent associated word pair that appear next to key term in the text corpus of that domain.
In order to specify the associated words that can appear in the mind next to each key term in the target domain, the model scan the text corpus looking for the nearest context word around each key term, and breaks the text into a windows of size 2 words, a key term plus the next or previous word.

Table (1) show the context windows for the key term "ضغط دم" and their next or previous context words in the text SemEval2016 training dataset 16, and Figure 3.Demonstrate the view of the semantics for the key term "ضغط دم" and shows thedirections ofsimilarity and association relations.

Due to the asymmetric property in association relation i.e. the relation between the word pairs "ضغط دم"and " دم ضغط " are not equivalent [8], our model process each window using bi-gram tokenization to preserves this essential property, then we end up with a list of associated word pairs that represent the special view of semantic constraints in the target domain. To eliminate weak association relation in the list we drop less than 5 frequent word pairs. Table (2) shows examples of word pairs in the extracted list of the special view of semantic constraints and their frequencies in the Arabic medical domain.

Table 1: Previous and Next Words for the Term "ضغط دم"

Next	*Key Term*	*Previous*
.....	ضغط دم	ارتفاع
		انخفاض
		قياس
		هبوط
مرتفع		.....
طبيعي		
منخفض		
يتراوح		

Figure 3: Similarity andAssociationRelationsDirections

Table 2: A List of Arabic Medical Domain View of Semantic Constraints and their Frequencies

Freq.	*Group(1)*	*Freq.*	*Group(2)*	*Freq.*	*Group(3)*	*Freq.*	*Group(4)*
1661	**ضغط** دم	661	**سكر** دم	396	مضادات **اكتئاب**	637	**التهاب** كبد
1257	ارتفاع **ضغط**	262	مستوي **سكر**	218	علاج **اكتئاب**	410	**التهاب** مفاصل
81	**ضغط**سكر	145	ارتفاع **سكر**	201	**اكتئاب** نفسي	298	**التهاب** مسالك
65	**دم** مرتفع	91	**سكر** ضغط	145	قلق **اكتئاب**	172	**التهاب** مزمن
28	**دم** منخفض	37	انخفاض **سكر**	48	اعراض **اكتئاب**	116	مضادات**التهاب**

5. Adapting Word Embedding

Equalize Our proposed adaptation method is based on an extracted view of semantic constraints of the target domain to adjust the semantic representation of a concrete pre-trained embedding, in order to obtain a high quality of domain-specific embedding.

This task can be performed via tightening the view of semantic in the pre-trained embedding to close fit to the view of semantic of the target-specific domain, which can allow pre-trained embedding to be used efficiently for any of the various specific domains.

For this task, our model deploys the attract part of the specialization method [11] to tighten the association relation between the word vectors of each word pairs in the list of special views of semantic constraints, See Figure 4a.andFigure 4b. The main advantages of the specialization method is that it implements standard L2 regularization to

retains the strength of the initial distributional of vector representation, and also it uses the negative samples to limit the update process to weakest relation only, negative samples are the nearest cosine similar vector pairs to the target word vector pairs correspond to word pairs in a sub-list from the list of the special view of semantic constraints called mini-batch [11].

The model cost function has two terms:
$C(B_{ass}, T_{ass}) = S(B_{ass}, T_{ass}) + R(B_{ass})$
Where:

- B_{ass} is the mini-batch, a sub-list of size K for word vector pairs $(\boldsymbol{x}_l^i, \boldsymbol{x}_r^i)$ correspond to word pairs (x_l^i, x_r^i) in the list of target view of semantic constraints, where i=1to k.
- T_{ass} is the negative example of word vectors pairs for mini-batch Bass.

The first term, see equation (1) tightens the words pairs using the target view of semantic constraints list:

$$S(B_{ass}, T_{ass}) = \sum_{i=1}^{k} [T(\delta_{ass} + \boldsymbol{x}_l^i \boldsymbol{t}_l^i - \boldsymbol{x}_l^i \boldsymbol{x}_r^i) + T\delta ass + \boldsymbol{x}ri\boldsymbol{t}ri - \boldsymbol{x}li\boldsymbol{x}ri \quad (1)$$

Where:

- $\boldsymbol{x}_l^i, \boldsymbol{x}_r^i$ is a target word vector pair that correspond to the word pair in the list of the target view of semantic constraints.
- $\boldsymbol{t}_l^i$, $\boldsymbol{t}_r^i$ is the negative example, which is the closest word vector pair to the target word vector pair.
- δ_{ass} is the association margin which determines how much the associated vectors should be tightened.

The T(x) = max (0, x) is the hinge loss function.

The second term, see equation (2) retains the initial distributional representation:

$$R(B_{ass}) = \sum_{x_i \in (B_{ass})}^{k} \lambda_{reg} \|\hat{x}_i - x_i\|_2 \quad (2)$$

Where:
λ_{reg} is the L2 regularization constant.
$\hat{x}_i$is the original word vector representation for word x_i.

Figure 4a.andFigure 4b.show the visualization for word embedding local trained on the Arabic medical domain before and after adapting using our method and the special view constraints for the medical domain in (Table 2).

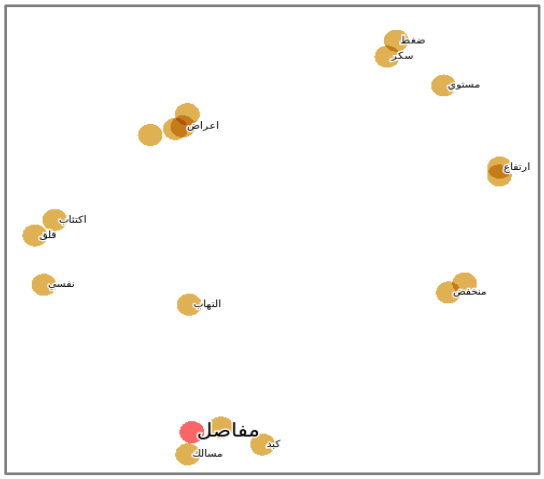

Figure 4a: Embedding Before Adaption

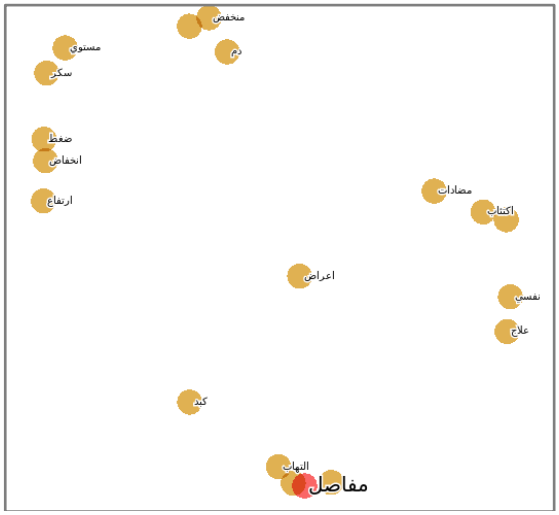

Figure 4b: Embedding After Adaption

6. Experiments

6.1 Task and Dataset

Experiments are conducted on SemEval-2016 Task-3 Arabic subtask-D [17]. The task is a re-ranking task for IR-based Community Question Answering (CQA) in Arabic medical domain, and the goal is to re-rank the candidate answers according to their relevancy to the given new question [17], for this task SemEval-2016 provides a dataset for training, development, and test, all candidate answers in the datasets are annotated to direct, relevant or irrelevant answers as shown in the table (3). The development data consist of 250 new user questions and a set of 30 candidate answers in a form of question-answer pairs for each new question.

Table 3: Training and Development Dataset Statistics

Training Dataset		Development Dataset	
Original Questions:		**Original Questions:**	
- TOTAL:	1,031	- TOTAL:	250
Retrieved Question-Answer Pairs:		**Retrieved Question-AnswerPairs:**-TOTAL:	7,384
- TOTAL:	30,411		
- Direct:	917	- Direct:	70
- Relevant:	17,412	- Relevant:	1,446
- Irrelevant:	12,082	- Irrelevant:	5,868

6.2 Proposed Method for CQA

Our method for the task mentioned above is mainly based on semantic resources and query processing techniques, the aim is to obtain a better understanding for the user question intent and then ranking the direct and relevant answers above irrelevant ones for each target question in the development dataset, The main part of the proposed method involves:

6.2.1Query Segmentation and Relaxation

Due to the length of user query in the medical domain, which is most often longer compared to query length in IR, each new question in the development dataset is segmented using the medical domain dictionary and then relaxed to the key medical terms plus one term left and/or right. The table (4) shows an example of a single new user question before

& after the segmentation and relaxation process. This sub-process achieved a 40.82 MAP score as shown in (Table 6).

Table 4: Segmentation and Relaxation Process

User Query	*Segmentation &Relaxation*
ثبت شكل قاطع لولب يوضع داخل رحم منع حمل الميرنا تسبب مرض سرطان	Before
رحم "منع حمل" تسبب "مرض سرطان"	After

6.2.2Query expansions based on an adapted word embedding

In the part syn-sets are extracted from an adapted version of word embedding using the view of semantic of the target domain using cosine distance measurement, see the default line in Table (5) which shows an extracted syn-set for the term "يعاني ".

6.2.3 Concept aware Query Expansion

Embedding based query expansion techniques generate syn-sets based on cosine similarity only, these syn-sets are fixed and used to expand all user questions while they have various concepts. The goal is to improve the quality of query expansion by generating syn-set regarding the concept of the current question. For CQA the set of candidate answers retrieved by the IR for each new user question also known as question thread represent the concept for that question, so for a given question we adopt its question thread to tune the expansion set for each term in that question, and if a term appears in several questions we can obtain a various expansion set for that term according to the concept of each question. The idea to eliminate each term in the syn-set that not belongs to the concept of the current question. Table (5) shows the syn-set for the term" يعاني " using cosine distance and concept-aware method for question number 177 and 176 in development dataset. This sub-process achieved a MAP score of 41.24 as shown in (Table 6).

Table 5: Default and Concept Aware Query Expansion

Syn-set size	*Syn-set "**يعاني**"*	*Syn-set Extraction Method*
10	تعاني,عاني,يشكو,يعانون,مصابا,اصيب,يشتكي,مصابون,يصاب,اعاني	Default
9	تعاني,عاني,يشكو,يعانون,مصابا,اصيب,مصابون,يصاب,اعاني	Question 177
3	تعاني,يصاب,اعاني	Question 176

6.3 Experimental Setup

6.3.1Data Pre-processing:

We follow a common text pre-processing pipeline that involves removing non-letters, numbers and normalizing a different form of the same character to a single form such as normalize آ,إ, أ to ا. Then we use MADAMIRA tool [24] for tokenization. This pipeline is adopted for both training and development dataset.

6.3.2Word Embedding:

To see how our adapting method improves the performance of embedding based query expansion for SemEval-2016 Task-3 Arabic subtask-D, A set of word embeddings are prepared which include:

- Locally trained embedding trained on SemEval-2016 Task-3 Arabic subtask-D training dataset [17], using the Eigen words algorithm [25], with the default parameters.
- Fine-tuned version of locally trained embedding, tuned using the local view constraints extracted from SemEval-2016 Task-3 Arabic subtask-D training dataset [17].
- An adapted version of universal pre-trained embedding for Arabic from FastText skip-Gram model [7], embedding is adapted using the local view constraints extracted from SemEval-2016 Task-3 Arabic subtask-D training dataset[17].

6.3.3Indexing and Re-ranking

Solr search engine tools are used to index the 30 QA-pair in question thread for each of 250 new questions in the development dataset in Semeval2016 Task3-D, expand all questions terms at query time and then re-rank its candidate answers.

Other tasks like query segmentation, relaxation, andconcept-aware expansion are performed using python code.

7. Results and Discussions

The results are evaluated using semeval2016 scorer tools [17]. The evaluation is based on the Mean Average Precision (MAP) score of the official task score. The results presented in Table (6) shows the progress in sub-methods performance. However,the performance is decreased in thesub-method that based on an adapted version of locally trained embedding which, adapted using Specialization method[11] and PPDB [20] semantic lexicon constraints, because the specialization method requires high quality of word embedding [11] which is hard to be found in locally trained embedding.

In (Table 7) compare an adapted version of locally trained embedding and universal pre-trained embedding, both adapted using our proposed model based on the view of semantic of the medical domain.The adapted universalpre-trained embedding with score 41.73outperforms the adapted local trained embedding with score 41.36.

Table (8) shows that our proposed method outperforms the state of the art method Word and context-word embedding Averaging 5.

Table 6: The Proposed Method phases Performance

Methods	*MAP Score*
Baseline	29.79
Query Segmentation(QS)	39.14
Query Relaxation (QR) + (QS)	40.82
Embedding based Query Expansion + (QR) + (QS)	**41.24**
Query Expansion based on fine-tuned embedding using PPDBsemantic constraints + (QR) + (QS)	40.38

Table 7: Local Trained and Universal Pre-trained EmbeddingAdapted Version Performance

Adaptation Method	*Embedding*	*MAP score*
The view of semantics	Local Trained	41.36
	Universal pre-trained	**41.73**

Table 8: The Performance of our Method and State of the Art Adapting Model

Adapting Model	*MAP score*
Word context-word Averaging	41.43
The View of Semantics	41.73

8. Conclusion

We have developed a novel method for adapting concrete universal pre-trained word embedding based on the view of semantic relation extracted from a text corpus and a dictionary of the target domain. And also, for the task of CQA, we have proposed a query relaxation method to relax the long user query to short query preserving the key terms of the question. Additionally, we also have proposed a technique that generates concept aware query expansion sets based on the thread of the target question.The results of the conducted experiments have shown that the performance of an adapted pre-trained word embedding outperformed the locally trained embedding in the task of CQA for Arabic. The results reveal the importance of handling similarity and association relations jointly in adopting an off the shelf word embedding in various specific domains.

References

[1] Devlin, J., et al., Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
[2] Ruder, S., Neural Transfer Learning for Natural Language Processing. 2019, NATIONAL UNIVERSITY OF IRELAND, GALWAY.
[3] Sarma, P.K., Y. Liang, and W.A. Sethares, Domain adapted word embeddings for improved sentiment classification. arXiv preprint arXiv:1805.04576, 2018.
[4] Mikolov, T., et al. Distributed representations of words and phrases and their compositionality. in Advances in neural information processing systems. 2013.
[5] Mikolov, T., et al., Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.
[6] Pennington, J., R. Socher, and C. Manning. Glove: Global vectors for word representation. in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.
[7] Bojanowski, P., et al., Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 2017. 5: p. 135-146.
[8] Asr, F.T., R. Zinkov, and M. Jones. Querying word embeddings for similarity and relatedness. in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018.
[9] Diaz, F., B. Mitra, and N. Craswell, Query expansion with locally-trained word embeddings. arXiv preprint arXiv:1605.07891, 2016.
[10] Mrkšić, N., et al., Counter-fitting word vectors to linguistic constraints. arXiv preprint arXiv:1603.00892, 2016.
[11] Mrkšić, N., et al., Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. Transactions of the association for Computational Linguistics, 2017. 5: p. 309-324.
[12] Yin, W. and H. Schütze. Learning word meta-embeddings. in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016.
[13] Joseph, K. and K.M. Carley. Relating semantic similarity and semantic association to how humans label other people. in Proceedings of the First Workshop on NLP and Computational Social Science. 2016.
[14] Wieting, J., et al., From paraphrase database to compositional paraphrase model and back. Transactions of the Association for Computational Linguistics, 2015. 3: p. 345-358.
[15] Jo, H., Deep Extrofitting: Specialization and Generalization of Expansional Retrofitting Word Vectors using Semantic Lexicons. arXiv preprint arXiv:1808.07337, 2018.
[16] Faruqui, M. and C. Dyer, Non-distributional word vector representations. arXiv preprint arXiv:1506.05230, 2015.
[17] AlessandroMoschitti, P.L., et al., Semeval-2016 task 3: Community question answering. Proceedings of SemEval, 2016: p. 525-545.
[18] Abadi, M., et al., Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467, 2016.
[19] Miller, G., WordNet: A Lexical Database for English Communications of the ACM Vol. 38. 1995.
[20] Ganitkevitch, J., B. Van Durme, and C. Callison-Burch. PPDB: The paraphrase database. in Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013.
[21] Navigli, R. and S.P. Ponzetto, BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 2012. 193: p. 217-250.
[22] Yousif, H. and R. Mohsen, Semantic-based Arabic Question Answering: Core and Recent Techniques. International Journal of u- and e-Service, Science and Technology 2Vol.10, no. 1, 2017, p.201-214.
[23] Pakhomov, S., et al. Semantic similarity and relatedness between clinical terms: an experimental study. in AMIA annual symposium proceedings. 2010. American Medical Informatics Association.
[24] Pasha, A., et al. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. in LREC. 2014.
[25] Dhillon, P.S., D.P. Foster, and L.H. Ungar, Eigenwords: Spectral word embeddings. The Journal of Machine Learning Research, 2015. 16(1): p. 3035-3078.