

A Contextual Semantic-Based Approach for Domain-Centric Lexicon Expansion

Muhammad Abulaish¹, Mohd Fazil¹, and Tarique Anwar^{2,3}

¹ South Asian University, New Delhi, India

{abulaish@sau.ac.in, mohdfazil.jmi@gmail.com}

² Macquarie University, Sydney, Australia {tarique.anwar@mq.edu.au}

³ CSIRO Data61, Sydney, Australia

Abstract. This paper presents a contextual semantic-based approach for expansion of an initial lexicon containing domain-centric seed words. Starting with a small lexicon containing some domain-centric seed words, the proposed approach models text corpus as a weighted word-graph, where the initial weight of a node (word) represents the contextual semantic-based association between the node and the target domain, and the weight of an edge represents the co-occurrence frequency of the respective nodes. The semantic-based association between a node and the target domain is calculated as a function of three contextual semantic-based association metrics. Thereafter, a random walk-based modified PageRank algorithm is applied on the weighted graph to rank and select the most relevant terms for domain-centric lexicon expansion. The proposed approach is evaluated over five datasets, and found to perform significantly better than three baselines and three state-of-the-art approaches.

Keywords: Text mining · Keyword extraction · Lexicon expansion · Contextual similarity

1 Introduction

Extraction of keywords or keyphrases from large text corpora is an important task in many text information processing applications, in which important and relevant words are extracted from the corpora. Such words are generally related to all the different domains of interest. Research on this problem in the past few decades has resulted into a rich literature [4]. While keywords are relevant to multiple domains of interest, they are not much effective in highlighting some specific domains. Lexicons, on the other hand, are able to effectively conceptualize one particular domain with relevant words from the text corpus. Lexicon-based approaches are highly effective in many applications, such as spam email classification, abusive language detection, sentiment analysis, and emotion mining. Although there exists many works on lexicon generation in the literature, they predominantly ignore the contextual semantics. It makes them ineffective over online social networks (OSN) data. Moreover, most of the existing lexicons are

generally curated through crowd-annotation [7,10], which is a time-consuming and tedious task. There exists some well-established benchmark lexicons such as Hatebase⁴, SocialSent⁵. but still there is a lack of sufficient lexicons to cover every domain of interest. For example, there is no such lexicon of radical words used by different extremist groups in the South Asian region. Through human efforts, one can possibly identify only a limited number of radical words such as *kashmirfreedom*, *gazwaehind*, *khalistan*. It is not feasible to manually identify all other contextual words that are used by such extremist groups. Therefore, automated lexicon expansion from a given initial lexicon of few seed words, is an important research problem.

There exists some works in the direction of domain-centric lexicon expansion from a text corpus, most of which use the concepts of contrasting corpora and graph-based approaches [8,3,9]. Sarna et al. [9] utilized an initial lexicon of seed words using a statistical significance analysis-based approach for its expansion. However, it ignores the contextual semantic of corpus words with the seed words. Overall, the existing works suffer from three major limitations. Firstly, most of the existing approaches are based on simple statistical measures like frequency count and co-occurrence count ignoring the contextual semantics between the terms. Secondly, to the best of our knowledge, all existing approaches except [9] are for lexicon generation rather than expansion. Finally, no approach exists that utilizes the strengths of both the contrasting-corpora and graph-based approaches incorporating contextual semantics towards initial lexicon expansion over the OSN data. To this end, this paper utilizes the advantages of both the statistics of contrasting-domain corpora and contextual semantics of latest word vector representation for domain-centric lexicon expansion. Further, proposed approach exploits an initial lexicon of few seed terms to bias the initial contextual semantic-based scores of corpus-words towards the target domain.

2 Proposed Approach

2.1 Candidate Words Extraction

The selection of content-bearing candidate words from the corpus is an important step of the lexicon expansion process. OSNs are a conversation platform where users generally use an informal and noisy language. Therefore, firstly, uninformative symbols and special characters like “@”, “#”, “RT” are filtered out from tweets, which are further converted to lower case to avoid ambiguity between words. The filtered tweets are further passed to a part-of-speech tagger to find *noun* and *adjective* phrases [5], which are generally important words in user-generated contents conceptualizing the text corpus. Finally, identified *noun* and *adjective* phrases are lemmatized to construct the set of potential candidate words for graph modeling.

⁴ <https://hatebase.org/>

⁵ <https://nlp.stanford.edu/projects/socialsent/>

2.2 Contextual Semantic-Based Graph Construction

The candidate words are modeled as a word co-occurrence graph $G = \langle W, E \rangle$, where W is the set of nodes representing the candidate words and E is the set of links connecting the nodes (words). Further, we compute the initial vertex score of each word representing the contextual semantic-based association between the word and target domain, and edge weight is assigned to show the co-occurrence frequency between every pair of words. An edge between a pair of words is created only when they have co-occurred in at least one document of the corpus (in our case, it is a tweet).

Vertex Relevance Score The initial weight assigned to a node (word) $w \in W$ is based on three association measures – (i) contextual semantic-based similarity of w with S , (ii) domain relevance of w with respect to a set of contrasting corpora, and finally (iii) occurrence-probability of w with seed words.

Embedding-Based Semantic Similarity The semantic similarity of w with the seed words of S is based on numeric vector representation of words. In the existing literature, several neural network-based methods have been presented to train the low dimensional numeric vector representation of words. In such an approach, Mikolov et al. [6] presented a computationally efficient and widely accepted approach to learn the word representation from unlabeled corpus using two models – (i) a continuous bag of words (CBOW) representation model and (ii) skip-gram model. In the proposed approach, we use the *CBOW* model to train a word embedding model that maps each word of the corpus into a low-dimensional vector in a vector space of latent concepts. Thereafter, semantic similarity between the lexicon of seed words and each word in the graph is computed based on the trained word-embedding vectors. The contextual semantic-based similarity of each word of G is the average of cosine similarity of the word with each seed word of S as given in equation 1, where e_w and e_s represent the embedding vectors of w and $s \in S$ respectively.

$$\mathcal{S}(w) = \frac{\sum_{s \in S} \text{Cos}(e_w, e_s)}{|S|} \quad (1)$$

Domain Relevance In the proposed approach, domain relevance of a word w is defined as the ratio of the occurrence probability of w in domain-specific corpus to the average of its occurrence probability in the contrasting corpora. If the domain-specific corpus is D and contrasting corpora C , then domain relevance $\mathcal{D}(w)$ of a word w is defined as given in equation 2, where P_w^D represents the occurrence probability of w in D and P_w^C represents the average of the occurrence probability of w in C .

$$\mathcal{D}(w) = \frac{P_w^D}{P_w^C} = \frac{tf_w^D / N^D}{\sum_{c \in C} \frac{tf_w^c}{N^c} / |C|} \quad (2)$$

Co-occurrence-Based Contextual Proximity The frequent occurrence of a word with seed words reflects its contextual proximity with seed words. Therefore, we define a metric called co-occurrence-based contextual proximity, \mathcal{P} , to capture the co-occurrence of a word with seed words. For a word w , it is the average of conditional probability of w with each $s \in S$ as defined using equation 3, where $p(s/w)$ represents the conditional probability of s given that w has already occurred.

$$\mathcal{P}(w) = \left(\sum_{s \in S} p(s/w) \right) / |S| \quad (3)$$

Finally, vertex score $\mathcal{V}(w)$ of w is defined as given in equation 4

$$\mathcal{V}(w) = (\mathcal{S}(w) + \mathcal{D}(w) + \mathcal{P}(w)) / 3 \quad (4)$$

Edge Score This section captures the contextual semantic-aware association between every pair of words $(w_i, w_j) \in G$ to create edges between them. We define the edge weight between a pair of words (w_i, w_j) of G as the number of tweets in which they co-occur regardless of any window size to incorporate tweet-level context. It is defined using equation 5, where $I_t(w_i, w_j)$ is the identify function which is one when both the words occur in a tweet t otherwise zero as given using equation 6.

$$\mathcal{E}(w_i, w_j) = \sum_{t \in D} I_t(w_i, w_j) \quad (5)$$

$$I_t(w_i, w_j) = \begin{cases} 1 & \text{if } w_i \in t \text{ and } w_j \in t \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Finally, we normalize the edges weight using equation 7, where \mathcal{E}_{max} represents the weight of the edge with the highest value.

$$\mathcal{E}(w_i, w_j) = \frac{\mathcal{E}(w_i, w_j)}{\mathcal{E}_{max}} \quad (7)$$

$$\mathcal{V}'(w_i) = \frac{1-d}{N} + \sum_{w'_i \in Adj(w_i)} C * \frac{\mathcal{E}(w'_i, w_i) * \mathcal{V}(w'_i)}{|Adj(w'_i)|} \quad (8)$$

2.3 Words Ranking and Lexicon Expansion

In **PageRank**, initial weight of nodes follows uniform distribution with an equal weight of 1. Thus, every node has equal probability of random jump to other nodes of the graph. On contrast, in the proposed approach, weights on nodes follow a non-uniform distribution such that the nodes having higher contextual semantic with seed words are assigned higher weights emulating the personalized

PageRank. The non-uniform distribution of weights biases the computation towards certain nodes in the recursive procedure. It allows the nodes of the graph to spread their importance to other nodes depending on their weights. This spread of a node score is also affected by the weights of adjacent edges such that the flow of weight will be higher between two strongly connected nodes. Therefore, the final weight of a node is not only based on its contextual semantic with the seed words but also depends on the strength of co-occurrence. Finally, a modified **PageRank** [2] is applied on G to identify the most relevant words for expansion of the initial lexicon of seed words. In the modified **PageRank** algorithm, importance score of a word is updated using equation 8, where $\mathcal{V}'(w)$ represents the updated score of $w \in W$, d is a damping constant (0.85), C is a scaling constant (0.95), and N is the number of words (nodes) in the graph. The iterative procedure of score updation of each word is repeated until a stationary distribution of words score is reached. Thereafter, words are sorted based on their final scores and high ranked words are selected for lexicon expansion.

3 Experimental Setup and Results

This section presents a detailed description of datasets and embedding learning, evaluation results, and comparative analysis.

3.1 Dataset and Embedding Learning

The proposed approach is evaluated on five different domains of datasets prepared using two main datasets – D_1 and D_2 . The dataset D_1 is a benchmark dataset of 80000 tweets related to three categories of offensive languages – *hateful*, *spam*, and *abusive* [1] including *normal* tweets. We crawled 64963 tweets (remaining were suspended) and their related metadata information from the provided tweet-ids to construct D_1 and learn the 100-d word-embeddings using **Word2Vec** model. Thereafter, a random set of 1000 tweets, called D_h , D_s , and D_a respectively, each from *hateful*, *spam*, and *abusive* categories are selected to evaluate the proposed approach. Further, three sets of 286, 343 and 264 keywords are extracted from D_h , D_s and D_a , respectively using the **Natural Language Understanding** tool. Thereafter, three annotators are asked to rate the extracted keywords on a 11-point scale from 0 to 10, where 0 is assigned when annotator is 100% confident that keyword does not belong to a particular category and it is assigned 10 when annotator is 100% confident that keyword belongs to a particular category. Finally, average of the three rating scores is compared with 5 to create an annotated set of 76 hate words (A_h), 130 spam words (A_s), and 105 abusive words (A_a).

To further evaluate the proposed approach, another dataset D_2 is crawled during August 5, 2019 to August 28, 2019 using **Twitter** based on 14 radical keyphrases related to *Khalistan* and *Kashmir* movements. Thereafter, same procedure is repeated on D_2 as on D_1 to learn embedding vectors and generate the set of ground-truth keywords. As a result, we have annotated sets of 48 and 90 keywords for *Khalistan* and *Kashmir* related tweets represented as A_{kh} and A_{ka} ,

Table 1: A brief statistic of five datasets

Category	Benchmark dataset D_1					Crawled dataset D_2		
	Abusive	Hateful	Spam	Normal	Total	Khalistan	Kashmir	Total
Total tweets	12878	2740	9048	40297	64963	3888	560	4448
Evaluation tweets	1000	1000	1000	1000	4000	1000	560	1560

respectively. A brief statistic about the five evaluation datasets is given in Table 1.

Table 2: Performance evaluation results using an initial lexicon of 3 seed words

	Datasets with 1000 tweets					Datasets with 500 tweets				
	D_h	D_s	D_a	D_{kh}	D_{ka}	D_h	D_s	D_a	D_{kh}	D_{ka}
P@80	0.350	0.738	0.500	0.463	0.600	0.338	0.588	0.437	0.337	0.537
R@80	0.368	0.454	0.381	0.771	0.533	0.355	0.362	0.333	0.562	0.477
F@80	0.359	0.562	0.432	0.578	0.565	0.346	0.447	0.378	0.421	0.506

3.2 Evaluation Results

The proposed approach is evaluated using three standard evaluation metrics – *precision*, *recall*, and *f-score* at K . The domain relevance $\mathcal{D}(w)$ for each word w of G is computed using a single contrasting corpus D_c of 1000 normal tweets from D_1 . Table 2 presents the evaluation results of the proposed approach at $K = 80$ over the five datasets using S containing 3 seed words. This table shows that in terms of $P@80$, proposed approach exhibits lower performance over D_h because the ground-truth set A_h has a number of words like *nazi*, *muslim*, *crazy* which are contextually used in hateful tweets but they were labeled by the annotators as *non-hatred* words. Moreover, many words like *terrorism*, *russia*, *gay*, which are used as hatred words in certain contexts were not extracted by the NLU. Accordingly, they are missing from the annotated set of words. In terms of $P@80$, proposed approach performs best on D_s as shown in bold typeface in the third row of Table 2, whereas, in terms of $R@80$, it performs best over D_{kh} dataset as shown in the fourth row of table 2. It is because that D_{kh} has the least number of manually annotated keywords, thereby, increases the recall. Similarly, performance evaluation results over the five datasets of 500 tweets in each is shown in the last five columns of Table 2. On analysis of evaluation results over 1000 and 500 tweets from Table 2, it can be observed that the performance of the proposed approach goes down as we decrease the number of tweets in the evaluation datasets to 500. A comparative evaluation of performance of the purposed approach over different values of k over the five evaluation datasets of 1000 words is shown in Figure 1. It can be observed from this figure that as we select less number of top ranked keywords for lexicon expansion, precision increases sharply whereas recall shows downgrading pattern as expected.

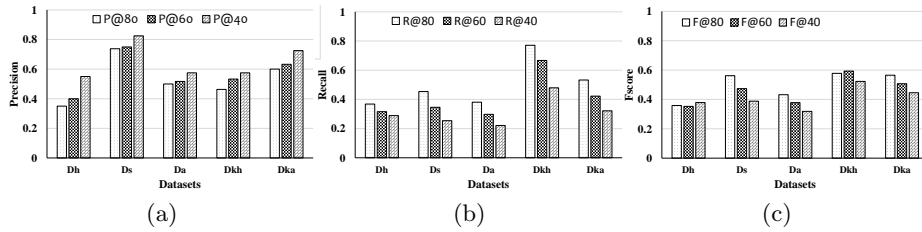

 Fig. 1: Performance evaluation results at different k values (80, 60 and 40)

Table 3: Comparative performance evaluation results

Approach	Datasets														
	D_h			D_s			D_a			D_{kh}			D_{ka}		
	P@80	R@80	F@80	P@80	R@80	F@80	P@80	R@80	F@80	P@80	R@80	F@80	P@80	R@80	F@80
Proposed Approach	0.350	0.368	0.359	0.738	0.454	0.562	0.500	0.381	0.432	0.463	0.771	0.578	0.600	0.533	0.565
Sarna and Bhatia [9]	0.244	0.128	0.168	0.318	0.121	0.175	0.209	0.086	0.122	0.250	0.102	0.144	0.286	0.106	0.155
Park et al. [8]	0.175	0.184	0.179	0.149	0.085	0.108	0.175	0.133	0.151	0.075	0.083	0.079	0.100	0.067	0.080
Kit and Liu [3]	0.163	0.171	0.167	0.350	0.215	0.267	0.025	0.019	0.022	0.075	0.125	0.094	0.125	0.111	0.118
tf	0.175	0.184	0.179	0.150	0.092	0.114	0.175	0.133	0.151	0.262	0.437	0.328	0.263	0.233	0.247
tf-idf	0.063	0.066	0.064	0.238	0.146	0.181	0.138	0.105	0.119	0.025	0.041	0.031	0.038	0.033	0.035
Embedding-based Similarity	0.075	0.079	0.077	0.113	0.069	0.086	0.325	0.248	0.281	0.262	0.438	0.328	0.388	0.344	0.364

3.3 Comparative Analysis

The proposed approach is compared with three baselines and three state-of-the-art approaches [9][8][3]. In the first baseline, we ranked and extracted the words based on their frequency count in text-corpus, whereas, second baseline extracts the top-ranked words based on their *tf-idf* value. Finally, in the third baseline, the embedding-based similarity of words with the lexicon of seed words is computed to extract the top ranked contextually semantic terms. Table 3 presents the performance evaluation results of the proposed approach in terms of all the three evaluation metrics for $K = 80$ in comparison to six approaches. It can be observed from this table that the proposed approach performs significantly better than all the comparison approaches. Among the three standard state-of-the-art approaches, [9] performs best though it shows poor performance in comparison to the proposed approach. Among the three baseline methods, words extracted using embedding-based similarity performs best over D_a , D_{kh} , and D_{ka} datasets whereas *tf-idf* based relevant word extraction performs worst. The *tf*-based relevant words extraction also shows good performance but not comparable to the proposed approach. The better results by embedding-based similarity also confirm the strength of the proposed approach, which uses contextual semantics based on the distributional representation of words as a measure of association between the corpus words and initial lexicon of seed words.

4 Conclusion

In this paper, we presented a contextual semantic-based approach utilizing the strengths of both the distributional word representation and contrasting-domain corpus for domain-specific lexicon expansion from text-corpus. We validated the performance of our approach by conducting experiments on five different Twitter datasets. Our approach performs significantly better in comparison to three baselines and three state-of-the-art approaches. The proposed approach is very useful for the domains in which the text corpus is not fixed, rather keeps incrementing with time.

Acknowledgment

The authors would like to thank the South Asian University, Delhi, for the financial support under the start-up research grant provided to the first author.

References

1. Founta, A.M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., Kourtellis, N.: Large scale crowdsourcing and characterization of twitter abusive behavior. In: Proceedings of the 12th Int'l Conference on Web and Social Media. pp. 491–500. AAAI, Palo Alto, California, USA (June 2018)
2. Hassan, S., Mihalcea, R., Banea, C.: Random-walk term weighting for improved text classification. In: Proceedings of the Int'l Conference on Semantic Computing. pp. 242–249. California, USA (September 2007)
3. Kit, C., Liu, X.: Measuring mono-word termhood by rank difference via corpus comparison. *Terminology* **14**(2), 204–229 (2008)
4. Matsuo, Y., Ishizuka, M.: Keyword extraction from a single document using word co-occurrence statistical information. In: Proceedings of the 16th Int'l Florida Artificial Intelligence Research Society Conference. pp. 392–396. AAAI, Florida, USA (May 2003)
5. Mihalcea, R., Tarau, P.: Textrank: Bringing order into text. In: Proceedings of the Int'l Conferences Empirical Methods in Natural Language Processing. pp. 404–411. ACL, Barcelona, Spain (July 2004)
6. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space **arXiv:1301.3781** (2013)
7. Mohammad, S., Turney, P.: Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* **29**(3), 436–465 (2013)
8. Park, Y., Patwardhan, S., Visweswariah, K., Gates, S.C.: An empirical analysis of word error rate and keyword error rate. In: Proceedings of the 9th Annual Conference of the International Speech Communication Association. pp. 270–273. Brisbane, Australia (September 2008)
9. Sarna, G., Bhatia, M.: A probabilistic approach to automatically extract new words from social media. In: Proceedings of the Int'l Conference on Advances in Social Networks Analysis and Mining. pp. 719–725. San Francisco, USA (August 2016)
10. Staiano, J., Guerini, R.M.: Depechemood: a lexicon for emotion analysis from crowd-annotated news. In: Proceedings of the 52nd Annual Meeting of the ACL. pp. 427–433. Maryland, USA (June 2014)