EAI Endorsed Transactions
on Scalable Information Systems Research Article **EAI.EU**

Semantic N-Gram Topic Modeling

Pooja Kherwa[1,*], Poonam Bansal[1]

[1]Maharaja Surajmal Institute of Technology, C-4 Janak Puri. GGSIPU. New Delhi-110058, India.

Abstract

In this paper a novel approach for effective topic modeling is presented. The approach is different from traditional vector space model-based topic modeling, where the Bag of Words (BOW) approach is followed. The novelty of our approach is that in phrase-based vector space, where critical measure like point wise mutual information (PMI) and log frequency based mutual dependency (LGMD)is applied and phrase's suitability for particular topic are calculated and best considerable semantic N-Gram phrases and terms are considered for further topic modeling. In this experiment the proposed semantic N-Gram topic modeling is compared with collocation Latent Dirichlet allocation(coll-LDA) and most appropriate state of the art topic modeling technique latent Dirichlet allocation (LDA). Results are evaluated and it was found that perplexity is drastically improved and found significant improvement in coherence score specifically for short text data set like movie reviews and political blogs.

Keywords. Topic Modeling, Latent Dirichlet Allocation, Point wise Mutual Information, Bag of words, Coherence, Perplexity.

Received on 16 November 2019, accepted on 10 February 2020, published on 11 February 2020

doi: 10.4108/eai.13-7-2018.163131

[*] Corresponding author. Email: poona281280@gmail.com

1. Introduction

Latent Dirichlet Allocation (LDA) is the most widely used probabilistic model in topic modeling literature. LDA is used to find the hidden semantic structure in the collection of documents. In LDA, the basic assumption is that each document is a distribution over topics and each topic is a distribution over fixed vocabulary [1]. It means we get two matrices of distributions from latent Dirichlet allocation output, first is distribution of terms over topic and second is distribution of topics over documents. The model assumes that each document consists of multiple topics and each document contain these topics in different proportion. It is common in information retrieval applications to treat documents as bag of words, ignoring any internal structure. Like other probabilistic models, the Latent Dirichlet allocation is also based on bag of words (Bow)assumption that treat each term of document collection independent from others. All the implicit semantic and syntactic structure is ignored in these language models. This bag of words assumption in vector space models simplifies the posterior inference in probabilistic models.

In this research, we are interested in topic model that capture the semantic order between words or dependency between adjacent words.

For e.g. The phrase 'NewYork' consists of two terms like New and York. Both terms are independent and can be divided and assigned to different topics and will lost its meaning as well as its context. So, we need a topic model, where the phrases consist of multiword can be considered as single term during pre-processing and assigned to correct topic accordingly.

In the topic modeling literature several extensions of LDA has been proposed that assign topics not only to individual words, but consider phrase as individual term as per semantic context and assign them to same topic [2][3]. A novel design for short text topic modelling, that encourages a regularization to encourage semantically related words to share the same topic assignment [4][5].A topic evolutionary graph from text ,which not only capture the main timeline, but also reveal correlation between related subtopics[6].In a similar work more coherent topics is presented known as Neural space topical coding (NSTC)[7][8]But most of the work consider only bigram based multiword phrase selection and another work is based on Adaptor Grammar framework (AG-Collc)[9][10]with a high time complexity. So, it is not feasible to use the model for large collection of text documents.

In this paper, a novel topic modeling approach based on semantic phrase extraction with more coherent topic detection is presented. In this collocation up to N-gram are extracted at pre-processing level from document collection, then methodology for phrase refinement is used before topic modeling. The novelty of approach lies in semantic extraction of collocation up to N-gram based on critical statistical measure like point wise mutual information (PMI) and log based mutual dependency, never done before. Here only contextual phrases crossing threshold of minimum score (PMI, LGMD*)* are extracted, renamed as semantic phrases, then latent Dirichlet allocation (LDA) is applied and significant improvement in performance of topic modeling in the terms of perplexity and coherence (with different level of cardinality in terms of topics) is achieved in topic modeling approaches.

The paper is organized as follows, in section 2, related work is discussed, in section 3, the proposed semantic N-Gram (SN-Gram) topic modeling framework is presented and in section 4, the proposed approach is validated with three datasets and finally the paper is concluded with future direction in section 5.

2. Background

There are numerous studies, where the bigram, idioms, multiword terms are investigated in topic models. HMMLDA (Hidden Markov Model Latent Dirichlet Allocation) was the first model to consider word dependency in the topic modeling architecture [11]. Bigram topic modeling (BTM) come into picture in 2006, with the concept of hierarchical Dirichlet language model, generates bigrams using word probabilities that are conditioned on the immediately preceding words[12][13]LDA-Collocation extends the bigram topic modeling with special variable to record the status of bigrams, this model is more realistic by giving flexibility to generate both unigrams and bigrams. Topical N-gram [14] proposed a model with significant contribution that can decide whether to form a bigram with two consecutive words according to nearby co-occurrence analysis. A semiformal(general)topic modeling framework based on semantic similarity is also good step towards non-probabilistic topic modeling [15]. Another work on phrase discovery topic modeling using hierarchical generative probabilistic model is able to relax the bag of word assumption using hierarchy of Pitman yor process [16] Top-Mine [17] is also a good tool that demonstrate scalability and interpretability compared to other phrase topic model but at the cost of high complexity of model. In 2010 Johnson establish the connection between LDA and probabilistic context free grammar to integrate the collocation and proper noun into the topic model [9][10]. In another work [18] challenges the assumption that the topic of N-gram is determined by the topics of composite words within the collocation. Unsupervised Multiview hierarchical embedding (UMHE)framework to sufficiently reveal the intrinsic topical knowledge in social events [19][20]. However, all these leads to increase in complexity and decrease in efficiency due to increase in the size of vocabulary of the model. Large vocabulary requires intensive computation. Collocation based topic modeling is classified into two categories, first is making probabilistic model with additional variables and generating distribution on them and second is extraction of N-gram in first stage of modeling and integration into Topic-Models.

3. Proposed Approach

Semantic N-Gram Topic Modeling (SNTM): The characteristics of latent Dirichlet allocation (LDA)algorithm has been used for finding the topics from large collection of documents or large text dataset [21]. In last two decades, many extensions of LDA like correlated topic modeling, dynamic topic modeling, supervised topic model, Author topic model, labelled topic model came into existence. All the methods hold the assumption of independence of topics in documents and based on bag of words (BoW) based vector space model. There are very few studies that have come forward to challenge this bag of word(Bow)models and to consider the importance of word order in effective topic modelling[18][22][15][3].Our proposed framework is semantic N-Gram topic modeling(SNTM), based on the importance of word order in sentences, in this novel approaches semantic order between words or dependency between adjacent words is captured at pre-processing level using special collocation extraction with statistical measure -pointwise mutual information and log based mutual dependency ,these collocation results in highly semantic N-Gram phrases, and value of N is chosen in the range of 1to5,so that semantically rich phrases should be included in vector space, and then topic modeling is applied for efficient and more coherent topic extraction.

Proposed approach can be divided into two sub modules:

1. Semantic N-Gram Phrase detection and refinement.
2. Topic modeling with Semantic phrase-based document term matrix.

The idea of semantic N-Gram Topic modeling is implemented with the following process.

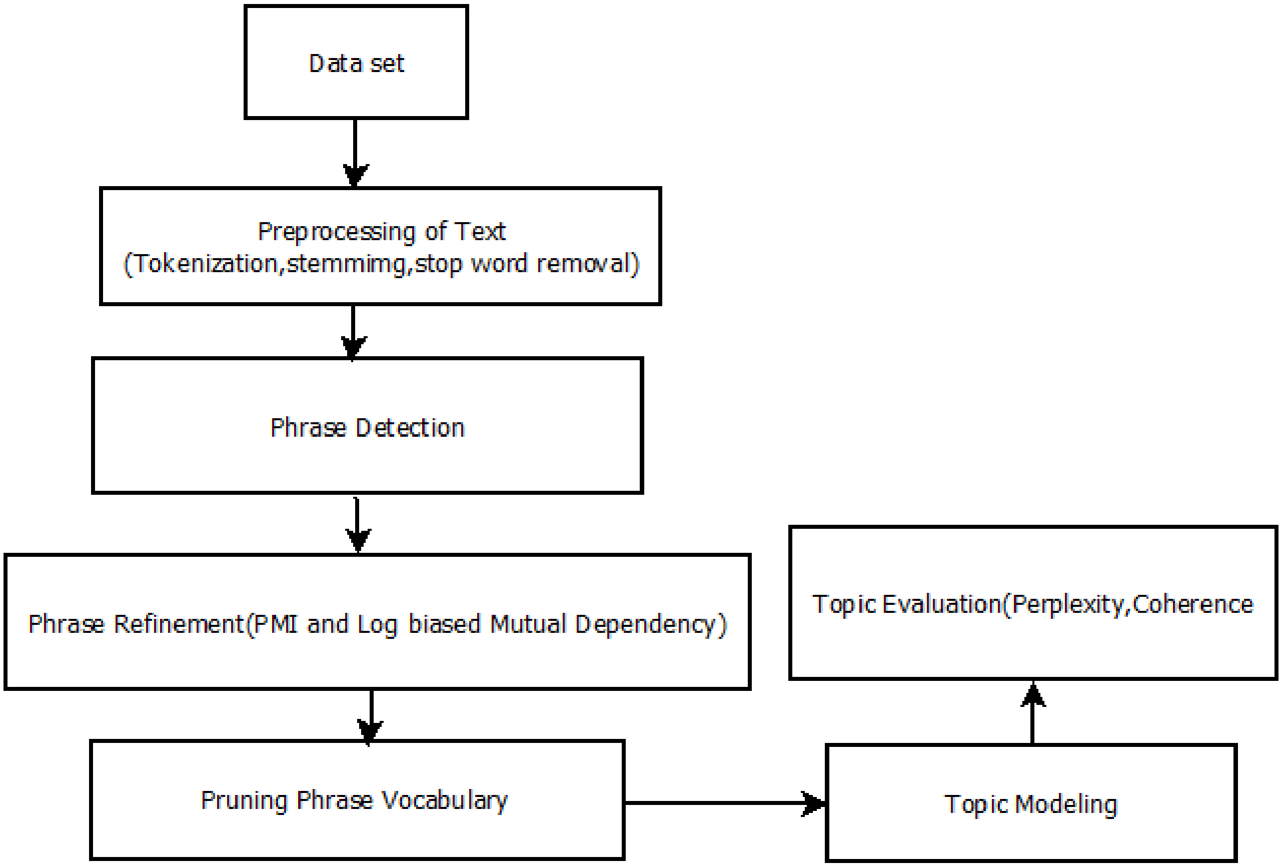
Data set
Preprocessing of Text (Tokenization,stemmimg,stop word removal)
Phrase Detection
Phrase Refinement(PMI and Log biased Mutual Dependency)
Pruning Phrase Vocabulary
Topic Modeling
Topic Evaluation(Perplexity,Coherence

Figure 1. Proposed Semantic N-Gram Topic Modeling (SNTM) framework

3.1 Semantic N-Gram Phrase Detection & Refinement

Semantic Phrases: Any sequence of words which when appears together give some meaning and when appears as individual word or token has no meaning is known as semantic phrases. In other words, it is a string of words which together behave like a single word.
The task of semantic N-Gram extraction consists of two-steps:

1. Phrase Extraction
2 Phrase Refinement

3.1.1 Phrase Extraction

This first phase of our proposed model implemented using text2vec package of R [23]. In this phase, first pre-processing of dataset like Stop-word removal, converting all vocabulary words to lowercase and considering only high frequency terms, all terms appear less than ten times are removed. Then all the collocation (Phrases) up to N-gram are extracted. These collocations are N-gram (consists of more than one word and up to N) whose probability of appearing together is greater than their individual frequencies and present the semantics of text in more meaningful way.

3.1.2 Phrase Refinement

The next step is the refinement of the list of N-Grams extracted in first phase Many of the extracted N-Grams are not meaningful. The aim of refinement is to extract only higher quality phrases from the list of extracted phrases and filter out the unnecessary collocation. For this purpose, some standard quality statistic measures are used. In the literature of natural language processing many pairwise significance measures for quality phrase detection are prevalent. In this research we choose:

3.1.2.1 Pointwise Mutual Information (PMI)

An effective method for semantic phrase extraction is the one that rank the actual phases above the other terms. The most widely used measure based on mutual information is the pointwise mutual information (PMI)considered in this research. PMI was introduced by Church and Hanks(1990)[3][24][13].PMI is a measure of how much the actual probability of a particular co-occurrence of events P(X,Y) differs from what we would expect it to be on the basis of the probabilities of the individual events and the assumption of independence P(X),P(Y)in some situation. In PMI low frequency events get high score, means infrequent word pairs tend to dominate the frequent events. So, they are ranked after the infrequent events.

This means that the PMI of perfectly correlated events words is higher when the combination is less frequent. So, in other words, PMI can be interpreting as a measure of independence rather than as a measure of correlation

$$I(W_{1,}W_{2}) = \log_{2}\frac{P(W_{1,}W_{2})}{P(W_{1}).P(W_{2})} \quad (1)$$

3.1.2.2 Log Biased Mutual Dependency (LBGD)

This is another collocation measure chosen for semantic phrase extraction because of its significant good performance compared to other collocation measure.

As already discussed, PMI gives preference to rare events as compare to frequent ones. So mutual dependency can be calculated in the phrases by subtracted from PMI the information that the whole event bears, which is self-information.

$$I\left(X\right) = -log\left(P\left(X\right)\right) \quad (2)$$

So mutual dependency (MD) can be defined as

$$D(W_{1,}W_{2}) = I(W_{1,}W_{2}) - I(W_{1}W_{2}) \quad (3)$$

$$D(W_{1,}W_{2}) = \log_{2}\frac{P^{2}(W_{1,}W_{2})}{P(W_{1}).P(W_{2})} \quad (4)$$

Mutual dependency will be maximized for perfectly dependent phrases or statistical confidence, it is suggested that slight bias towards frequency can be beneficial. So, a new measure known as log frequency biased MD can be defined as

$$D_{LF}(W_{1}W_{2}) = D(W_{1,}W_{2}) + \log_{2}P(W_{1}W_{2}) \quad (5)$$

In other words, it is combination of Mutual Dependency with T-score

3.2. Topic Modeling

In the second phase of proposed model, we use latent Dirichlet Allocation topic modeling algorithm on semantic phrases obtained in first phase of model.

Latent Dirichlet Allocation: Latent Dirichlet Allocation (LDA) is an approach based on definneti theorem [1] to capture significant inter as well as intra document statistical structure via the mixing distribution. LDA assumes that document arise from multiple topics. A topic is defined as distribution over a vocabulary. Corpus is associated with predefined number of topics k, and each document in corpus contain these topics with different proportion. Topic Modeling aims to learn these topics from data or corpus. LDA is based on hidden variable model, prevalent in machine learning from decades. Generative model does not consider order of words in producing documents, so purely based on bag of words (BOW) approach.

Topic modeling is a technique to automatically discover the topics from a collection of documents. The K topic distribution must be learned through statistical inference. In generative probabilistic modeling, the most widely used topic modeling algorithm is latent Dirichlet allocation (LDA). This algorithm defined generative process as a joint probability distribution over both observed and hidden variables. The process of learning the topic distribution is described through plate notation given in figure 2.

The distribution of latent variables in the given document is as shown in equation (6).

$$P(\theta,z|w,\alpha,\beta) = \frac{P(\theta,z,w|\alpha,\beta)}{P(w|\alpha,\beta)} \quad (6)$$

All the probabilistic topic modeling approaches has the potential to work in multilingual environment [25] The aim of Topic modeling is to automatically discover the topics from a collection of documents. The documents are observed, while the topic structure, the topics, per document topic distribution and per document per word topic assignment is hidden structure. The utility of topic modeling is coming from the property that the inferred hidden structure resembles the thematic structure of the collection.

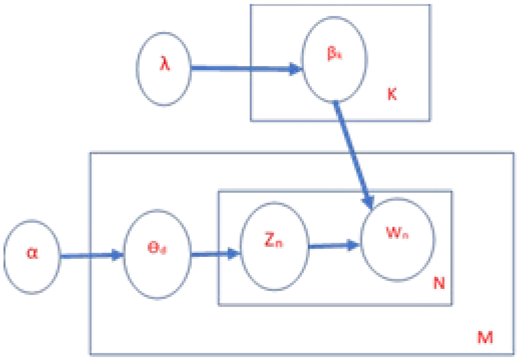

Figure 2: Latent Dirichlet Allocation Plate Notation [28]

3.3 Pseudocode SNTM

```
1.Take data set
2.Preprocess the data set by stop word removal, stemming and removing low frequency words.
3.Generate Phrases from 2 to N-Gram. (we can fix the size of N-Gram)
4.Select the best phrases for Semantic N-Gram selection using multiple statistics:
  (a) PMI (Point-wise Mutual Information)
  (b) LFMD (Log biased Mutual Dependency i.e. combination of t-score and mutual dependency (MD).
5.Construct document term matrix from phrase refined using statistical measures in (a) and (b).
6.Apply latent Dirichlet Allocation to newly constructed refined document term matrix.
7.Evaluate perplexity and coherence of semantic N-Gram Topic Modeling and compare result with Collocation Latent Dirichlet allocation (Coll-LDA) and Latent Dirichlet allocation (LDA).
```

4. Experimental Set Up

4.1. Data Set Used

In this experiment, we used three data set Movies review data set, Political blog dataset and ABC News headlines data set. Two data set are short text data set like Movie review data set is about movies reviews classified as positive and negative, and another is political blogs shared during USA presidential election in 2008. Third data set used is ABC news headlines dataset. This dataset contains news headlines in large text.

1. Movies Review

The labelled dataset consists of 5000 IMDB movie reviews, specially selected for sentiment analysis. The sentiment of the reviews is binary, meaning an IMDB rating < 5 results in a sentiment score of 0, and a rating >=7 has a sentiment score of 1. No individual movie has more than 30 review

2. Political-blog Data set

It is collection of blog-post collected from the CMU 2008 Political Blog Corpus. The blogposts were gathered from six different blogs American Thinker, Digby, Hot Air, Michelle Malkin, Think Progress, and Talking Points Memo. Each blog has its own particular political bent. Political blog data set is a collection of 13258 entries of 3 variables.

3.ABC News

This contains data of news headlines published over a period of 15 years. It sourced from the reputable Australian new source ABC (Australian Broadcasting Corp.it consists of 1076219 unique values with 1.10M*2(rows and column) with size of around 18 MB.

4.2 Pre-processing

For pre-processing of data set, a pre-processor function is designed which aims to remove all the numeric digits from dataset and also to convert all capital letters into lowercase for further processing. In next step a word tokenizer is also constructed and applied to the pre-processed input. This tokenizer will split the document into individual word level tokens.

A vocabulary-based document term matrix is created from all the terms. A Collocation function is applied to document term matrix to create long phrases of length up to 5. For efficient modelling we consider only those phrases which appears minimum 5 times, so that any accidental phrase should not become part of phrase vocabulary. After phrase identification, document phrase matrix is constructed for further processing

4.3 Topic Modeling using Semantic N-Gram Topic Modeling (SNTM)

The proposed approach is implemented with fixed number of topics for all the three data sets. The SNTM approach is constructed from the semantic vector space representation of text documents and compared with Coll-Lda (collocation based Latent Dirichlet Allocation) and latent Dirichlet Allocation (LDA)topic modeling approaches. For semantic N-Gram selection in SNTM the extracted phrases are not blindly used in topic modeling. In our approach some very efficient statistical measures are applied and best phrases with approved threshold are chosen and then topic modelling algorithms is applied.

4.4 Evaluation Framework

For evaluating the quality of inferred topics from SNTM Topic modeling the chosen measures are perplexity since it is the standard criterion of topic model's quality [26].

4.4.1 Perplexity

The most widely used approach in probabilistic model is to measure the log-likelihood of a held-out test set. For latent Dirichlet allocation a test set is the collection of unseen

documents Wd. and the latent model is described by the topic matrix Φ and the α hyperparameter for topic distribution of documents. Therefore, log likelihood is given by

$$L(w) = logP(W \mid \Theta, \alpha) = \sum_{d} \log P(Wd \mid \Theta, \alpha)$$

(8)

In this given topic Φ and the hyperparameter α for topic distribution ϴd of documents, higher loglikelihood implying a better model. In the language of topic modeling the term perplexity is used to evaluate topic model. It is inversely proportional to the loglikelihood.

Perplexity (test set w) = $\exp\{\dfrac{L(w)}{Countoftokens}\}$ (9)

Where L(w) is the loglikelihood of the unseen documents Wd, the lower the perplexity, the better the model.

4.4.2 Coherence

In the literature of topic modeling many measures based on word co-occurrence score of top terms of each individual topic has been used. Assessing coherence of topics is considered as the best way to give an insight on the interpretability of topics. Human topic ranking serves as the gold standard for coherence evaluation; however, they are expensive. For showing the goodness of our proposed model, in this experiment, we use a special case of topic coherence known as "mean PMI" (mean pointwise mutual information (that perform better than standard PMI (pointwise mutual information) in terms of correlation with respect to human topic ranking [27]. This Coherence metric uses pointwise mutual information (PMI) [27] for better interpretability of topics. PMI is calculated over the log of probabilities therefore it is negative. So, as the value of PMI approaches to 0, the topic coherence gets better.

5. Result & Evaluation

In this paper, we experiment with three data set and run the proposed efficient topic modeling SNTM on three data set. The first two data set are short text data set consists of Movie

reviews classified as positive or negative. and the second one is political blogs shared during presidential election in USA

in 2008. Third data set is about news headlines of ABC-News.

Movies review dataset has 5000 rows or entities of three variables. ABC News Headline dataset is a collection of 117965 objects of 2 variables, means 117965 headlines text with publication date.

Political blog data set is a collection of 13258 entries of 3 variables. Pre-processing of data set contain all necessary steps including stop word removal, replace capital letters to lower case, most importantly, we use a filter for vocabulary pruning and consider only those terms that appears at least ten times, and others are removed. After pre-processing of data sets the whole data set is converted into document term matrix.

5.1 Parameter Setting for Experiments

In this experiment after pre-processing a collocation model is constructed for learning phrases in dataset up to N Grams. For these collocation modeling number of iterations for all three datasets is taken 5. After collocation extraction the parameter for phrase refinement are taken as pointwise mutual information (PMI)set as 5 means phrases with PMI score greater than or equal to 5 are considered and other phrases are discarded. Also Log based mutual dependency (LBMD)score is set as (-25) to kept only meaningful phrases .One another important parameter in semantic N-Gram topic modeling is Collocation count to be considered so that unnecessary collocation can be avoided, it is assumed that if a phrase appear min ten times as phrase than it will be considered as meaningful phrase ,we take collocation count as 10.for example words "new" and "york" occurs 100 times together as "new_york" and each of words "new" and "york" occur 150 and 115 respectively. So intuitively there is a very high chance that "new_york" is good phrase candidate. In another case if we take a look at words "it", "is" it can happen that "it_is" occurs 500 times, but words "it" and "is" separately occur 15000 and 17000 times. Intuitively it is very unlikely that "it_is" represents good phrase. Other parameters for topic modeling are number of topics(K)and number of iterations. In this experiment for all three-dataset range of topics are taken from 5 to 35 and number of iterations are fixed at 10.

Experiments are evaluated using followings criteria including

5.2 Perplexity for different number of topics in Semantic N-Gram topic model, Coll- LDA and LDA

In movie review data set ,after pre-processing of data set, multi word phrases are extracted using ten iteration and finally converged at 6th iteration with total phrases of 2467.These phrases are again processed using (prune vocabulary)function with minimum term count min=10,it means phrases with term count value less than 10 are removed from the vocabulary, phrases with term count>10 are only retained in vocabulary list. The number of topics selected for modeling is in the range of 10-35 and number of iterations for both the topic modeling algorithms is fixed at 10. Parameter setting for all three data sets and results for all the three data set are shown in table 2.It is clearly visible in the table 2, that the perplexity of SNTM for all the selected data set is much lower than Collocation-Latent Dirichlet Allocation (Coll-LDA) and latent Dirichlet allocation (LDA) topic modeling. Low perplexity is better for topic modeling.

5.3 Coherence score with different number of topics

In this we choose Mean PMI (Point wise mutual information) as coherence metric, the motivation behind using Mean PMI is for better results than mutual information and pointwise mutual information for evaluating the topic modeling work. The topic modeling using Coll-LDA, LDA and Semantic N-Gram topic modeling (SNTM) approaches are evaluated using Mean-PMI for all the three data set discussed above using different number of topics ranges from 5 to 35. It is observed in experiment that coherence score (mean PMI) is also improved with our proposed approach of SNTM as shown in table 3. In the results of all three-dataset coherence score in “Movies review” dataset shows the highest coherence with our proposed SNTM topic modelling in comparisons with Coll-LDA and LDA. In other two dataset Poliblog dataset and ABC News-headlines SNTM approach gives best coherence score at topics range 20-35. At topic 5-15 Latent Dirichlet allocation and Coll-LDA perform better than our proposed approach. In ABC New headlines dataset SNTM gave good coherence score in topics 5-35 except at topic 10 and 20, where Coll-LDA gave best coherence score. Overall in most of the topic range our proposed approach SNTM give significant improved results.

5.4 Perplexity and number of iterations

As we experiment in last two section with different number of topics and fixed the number of iteration and checked the corresponding perplexity and coherence. Here we also modify the number of iteration and check its effects on perplexity, we take iteration from 10 to 50 and fixed the number of topics at 10 in movies review data set. Results shows good performance of semantic N-Gram Topic modeling in comparison with collocation latent Dirichlet allocation (Coll-lda) and traditional Latent Dirichlet topic modeling by providing lowest range of perplexity values.

5.5 Top Terms for Topics in Three Topic Models

To understand the topics generated from topic model, the most prevalent way to evaluate and understand topic modeling algorithms is the top term of individual topics. These top terms further used to compare different topic models, and calculate standard measure like coherence, log-likelihood for topic modeling algorithms. In the given table 3, we shown sample topics for ABC News Headlines using K=10, and number of iterations is 20, and topic considered are topic 1, topic 5, topic 10. we considered top ten terms for three topic models.

Table 1. Perplexity score of LDA topic model and Semantic N-Gram topic modeling with different number of topics.

Data Set Used	Number of Topics(K)	Perplexity SNTM	Perplexity of LDA-Coll	Perplexity of LDA
Movies Review	10	1280	1584	1952
	15	1252	1560	1919
	20	1242	1539	1883

	25	1230	1523	1858
	30	1214	1512	1847
	35	1210	1505	1834
Poliblog Dataset	5	2014	2051	2028
	10	1938	1976	1958
	15	1903	1926	1907
	20	1870	1905	1881
	25	1849	1877	1862
	30	1837	1855	1847
	35	1824	1845	1833
	40	1818	1834	1823
ABC News Headlines Data Set	10	258.133	263.66	257.25
	15	228.59	235.99	239.79
	20	218.165	219.40	219.56
	25	209.082	204.65	204.98
	30	203.85	200.16	202.56
	35	188.96	196.73	189.51
	40	187.22	181.34	193.26

Table 2. Coherence score of LDA topic model and Semantic N-Gram topic modeling with different number of topics.

Data Set Used	**Number of Topics(K)**	**SNTM-Mean PMI**	**Coll LDA-Mean-PMI**	**LDA-Mean Pmi**
Movies Review	5	-0.53	-0.55	-0.62
	10	-0.60	-0.62	-0.68
	15	-0.51	-0.51	-0.56
	20	-0.49	-0.51	-0.51
	25	-0.60	-0.68	-0.67
	30	-0.62	-1.23	-0.63
	35	-0.48	-0.46	-0.73
Poli-blog Dataset	5	-15.83	**-13.78**	-20.58
	10	-20.49	-19.53	**-17.90**
	15	-22.08	**-21.86**	-22.64
	20	-20.15	-22.65	-24.22
	25	-24.16	-24.38	-26.48
	30	-25.85	-26.67	-27.68
	35	-25.72	-26.15	-26.61
ABC News Headlines Data Set	5	-7.46	-10.75	-8.06
	10	-12.78	**-9.23**	-12.28
	15	-12.48	**-10.59**	-11.57
	20	-12.05	-12.17	-12.56
	25	-14.42	-14.87	-15.90
	30	-11.77	-13.99	-10.84
	35	-14.17	-15.24	-15.87

Table 3. Top Ten terms for three topic modelling techniques in ABC News Headlines Datase

Latent Dirichlet Allocation			Coll-LDA			Semantic N-Gram Topic Modeling		
Topic 1	Topic 3	Topic 5	Topic 1	Topic 3	Topic 5	Topic 1	Topic 3	Topic 5
missiles	final	offer	rejects	oil	second	Défense	record	missing
six	Brisbane	resolution	blues	Canberra	power	return	protesters	child
pledges	airs	France	offers	pair	cyclone	Hewitt	troops	power
vote	10	coast	day	bushfires	crean	banned	march	years
hopes	worlds	farm	worlds	remain	share	league	international	all
service	massive	state	Palestinian	yet	smoking	increase	Philippines	runs
destroys	tv	must	Powell	media	international	real	join	have
find	makes	cut	visit	investigate	house	course	against	season
rail	house	shire	win	suspect	south	injuries	sea	crows"
Drugs	Dubai	flood	meeting"	ban	home	Powell	must	use

6. Conclusion

In this paper an efficient approach called Semantic N-Gram topic modeling (SNTM) is presented with experimental results. It is different from prevalent collocation-based topic modeling in terms of efficiency and simplicity. In this instead of incorporating extra variables for inference in N-gram detection, our proposed approach emphasis on semantic N-gram extraction through statistical measures like pointwise mutual information (PMI) and log biased mutual dependency (combination of t-score and mutual dependency) and drop useless n-gram, those are below the considered threshold of statistical measures. After collocation refinement, latent Dirichlet allocation is applied and results are evaluated through standard evaluation measures of topic modeling like perplexity and coherence. It is clearly visible in the results that our proposed Semantic N-Gram Topic Modeling (SNTM) has depicted the improved performance in the evaluation measures perplexity. One important point to be noted is that collocation-based topic modeling approach gives best result in terms of coherence only in short text data set like "Movie Review" data set and "Poliblog" data set. In news headlines dataset "ABC News" the prevalent topic modeling approach latent Dirichlet algorithms (LDA) gives best results as compared to coll-LDA and SNTM topic modeling, may be these techniques creates unnecessary collocation in large text data set and generate meaning less phrases and reduced the coherence score.

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, volume 3: 993–1022.

[2] Hanna M. Wallach (2006). Topic Modeling: Beyond Bag of-Words. *Proceedings of the 23rd International Conference on Machine Learning*: 977–984.

[3] Jordan Boyd-Graber, David M. Blei, and Xiaojin Zhu. (2007). A Topic Model for Word Sense Disambiguation. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*: 1024–1033.

[4] Gao, W., Peng, M., Wang, H., Zhang, Y., Xie, Q., & Tian, G. (2019). Incorporating word embeddings into topic modeling of short text. *Knowledge and Information Systems*, 61(2), 1123-1145.

[5] Peng, M., Chen, D., Xie, Q., Zhang, Y., Wang, H., Hu, G., ... & Zhang, Y. (2018, November). Topic-net conversation model. *In International Conference on Web Information Systems Engineering* (pp. 483-496). Springer, Cham

[6] Gao, W., Peng, M., Wang, H., Zhang, Y., Han, W., Hu, G., & Xie, Q. (2019). Generation of topic evolution graphs from short text streams. *Neurocomputing.*

[7] Peng, M., Xie, Q., Zhang, Y., Wang, H., Zhang, X. J., Huang, J., & Tian, G. (2018, July). Neural sparse topical coding. *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 2332-2340).

[8] Peng, M., Xie, Q., Wang, H., Zhang, Y., & Tian, G. (2018). Bayesian sparse topical coding. *IEEE Transactions on Knowledge and Data Engineering*, 31(6), 1080-1093.

[9] Lindsey, R. V., Headden III, W. P., & Stipicevic, M. J. (2012). A phrase-discovering topic model using hierarchical pitman-yor processes. *In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*:214-222

[10] Mark Johnson M. (2010) PCFGs, Topic Models, Adaptor Grammars and Learning Topical Collocations and the Structure of Proper Names. *Proceedings of the 48thAnnual Meeting of the ACL*: 1148–1157

[11] Selivanov D (2017) text2vec: Vectorization. Retrieved-
from http://text2vec.org/vectorization.html

[12] Hanna M. Wallach (2006). Topic Modeling: Beyond Bag of-Words. *Proceedings of the 23rd International Conference on Machine Learning:* 977–984.

[13] Jey Han Lau, Timothy Baldwin, and David Newman. (2013) On Collocations and Topic Models. *ACM Transactions on Speech and Language Processing*,10(3): 1–14.

[14] Wei Hu, Nobuyuki Shimizu, Hiroshi Nakagawa, and Huanye Shenq. (2008). Modeling Chinese Document with Topical Word-Character Models. *Proceedings of the 22nd International Conference on Computational Linguistics*: 345–352

[15] Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. (2007) Topics in Semantic Representation. *Psychological Review*,114(2): 211–244.

[16] Kherwa, P., & Bansal, P. (2019). Empirical Evaluation of Inference Technique for Topic Models. *In Progress in Advanced Computing and Intelligent Engineering:*237-246.

[17] El-Kishky, A., Song, Y., Wang, C., Voss, C. R., & Han, J. (2014). Scalable topical phrase mining from text corpora. *Proceedings of the VLDB Endowment*, 8(3): 305-316.

[18] Viegas, F., Luiz, W., Gomes, C., Khatibi, A., Canuto, S., Mourão, F.,& Gonçalves, M. A. (2018) Semantically-Enhanced Topic Modeling. *In Proceedings of the 27th ACM International Conference on Information and Knowledge Management*: 893-902.

[19] Peng, M., Zhu, J., Wang, H., Li, X., Zhang, Y., Zhang, X., & Tian, G. (2018). Mining event-oriented topics in microblog stream with unsupervised multi view hierarchical embedding. *ACM Transactions on Knowledge Discovery from Data (TKDD*), 12(3), 1-26.

[20] Xie, Q., Peng, M., Huang, J., Wang, B., & Wang, H. (2019, July). Discriminative Regularization with Conditional Generative Adversarial Nets for Semi-Supervised Learning. *In 2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE

[21] T. Griffiths, M. Steyvers, D. Blei, and J. Tenenbaum. (2005) Integrating topics and syntax. *In Advances in Neural Information Processing Systems* 17.

[22] Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. (2007) Topics in Semantic Representation. *Psychological Review*,114(2): 211–244.

[23] Manning C. and Schütze H. (1999). Foundations of Statistical Natural Language Processing. *Cambridge: MIT Press*.

[24] De Finetti, B, (2017) Theory of probability: A critical introductory treatment, Vol. 6, John Wiley & Sons,2017.

[25] Ayadi, R., Maraoui, M., & Zrigui, M. (2014). Latent topic model for indexing arabic documents. *International Journal of Information Retrieval Research (IJIRR),* 4(2): 57-72.

[26] Ali Daud, Juanzi Li, Lizhu Zhou, Faqir Muhammad(2010).Knowledge discovery through directed probabilistic topic models: a survey. *Frontiers of Computer Science in China*, 4(2): 280–301

[27] Röder, M., Both, A., & Hinneburg, A. (2015) Exploring the space of topic coherence measures. *In Proceedings of the eighth ACM international conference on Web search and data minin*: 399-408

[28] Kherwa, P., & Bansal, P. (2020). Topic Modeling: A Comprehensive Review. *EAI Endorsed Transactions on Scalable Information Systems,* 7(24).

Pooja Kherwa is an assistant professor of Maharaja Surajmal Institute of Technology, New Delhi. She received her M. Tech in information Technology from Guru Govind Singh Indraprastha University; Dwarka, New Delhi in 2010.Currently she is pursuing her PhD from Guru Govind Singh Indraprastha University, Dwarka- New Delhi

Her research interest includes Topic Modeling, Sentiment Analysis, Machine Learning.

Dr. Poonam Bansal is a Professor of Maharaja Surajmal Institute of Technology. Also working as Deputy Director of Institute. She has received her PhD from Guru Govind Singh Indraprastha University, Dwarka, New Delhi in 2010.Her area of interest includes Speech recognition, Data Mining, Machine learning