Improving Coherence of Topic Based Aspect Clusters using Domain Knowledge

Kavita Asnani, Jyoti D. Pawar

Goa University,
Department of Computer Science and Technology,
India

{dcst.kavita, jdp}@unigoa.ac.in

Abstract. Web is loaded with opinion data belonging to multiple domains. Probabilistic topic models such as Probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA) have been popularly used to obtain thematic representations called topic-based aspects from the opinion data. These topic-based aspects are then clustered to obtain semantically related groups, by algorithms such as Automated Knowledge LDA (AKL). However, there are two main shortcomings with these algorithms namely the cluster of topics obtained sometimes lack coherence to accurately represent relevant aspects in the cluster and the popular or common words which are referred to as the generic topics are found to occur across clusters in different domains. In this paper we have used context domain knowledge from a publicly available lexical resource to increase the coherence of topic-based aspect clusters and discriminate domain-specific semantically relevant topical aspects from generic aspects shared across the domains. BabelNet was used as the lexical resource. The dataset comprised of product reviews from 36 product domains, containing 1000 reviews from each domain and 14 clusters per domain. Also, frequent topical aspects across topic clusters indicate occurrence of generic aspects. The average elimination of incoherent aspects was found to be 28.84%. The trend generated by UMass metric shows improved topic coherence and also better cluster quality is obtained as the average entropy without eliminated values was 0.876 and with elimination was 0.906.

Keywords. Topic-based aspect extraction, aspect filtering, aspect coherence, lexical resource BabelNet, context domain knowledge.

1 Introduction

In recent years, usage of social media for communication has increased enormously. The number of active users on social media has gone up exponentially and they express their opinions online. Therefore, most of the research on social media has concentrated on data which mostly occurs in text form and usually is more casual and context dependent. The commercial organizations and administrative users often look forward at the large collections of such opinion data to find useful patterns of significance and the related process is referred to as 'opining mining'.

Traditionally opinions can be expressed at document level, sentence level and aspect level [13]. To maximize the value from the opinions we need to process opinions at the fine grained level of granularity.

So we propose to work at the aspect level. In order to perform aspect discovery in social media context, there are several challenges that need to be addressed. Few of the prominent ones are listed below:

— Varied length messages: Social media text comprises of short and long length messages where the related semantic is highly dispersed in the content.

— Noisy: The chat style of social communication involves use of slang in the form of misspellings and ungrammatical words.

Computación y Sistemas, Vol. 22, No. 4, 2018, pp. 1403–1414
ISSN 1405-5546
doi: 10.13053/CyS-22-4-2401

— Lack of training data: Training may be necessary in supervised aspect discovery but practically there is lack of already existing pre-annotated data.

Fundamentally, the task of aspect based analysis comprises of identifying and extracting aspects of an entity in a domain on which opinions have been expressed [13]. We deal with extraction of relevant aspects automatically. With very large volume of data, useful aspects are often conveyed through latent semantics. Probabilistic Latent Semantic Analysis (pLSA) [10] and Latent Dirichlet Allocation(LDA) [3] are popularly recommended unsupervised topic modeling methods used for performing both extracting and grouping of related aspects [23]. However, they result in extracting some incoherent topics [4].

This is due to the occurrence of some irrelevant and polysemous terms in extraction [4]. [15] reported that models like pLSA and LDA do not provide coherent results, due to the objective function not correlating well with human judgments.

In addition, being unsupervised methods, these algorithms make no use of domain knowledge. [5, 7] proposed semi-supervised methods based on LDA for extracting aspects across the domains and expressed them as topic-based aspects. In [5], the topic-based aspects were clustered using Automated Knowledge LDA (AKL) and were accumulated in a common topic base. [5] used Gibbs sampling as an inference method. In an attempt to group semantically similar topic-based aspects for expressing coherence, k-mediod method was used in AKL to generate clusters.

However, as reported in [2] and [12] the clustering accuracy greatly improves by inputting domain knowledge. In [2] and [12] domain knowledge about relationship between words is used to force the assignments of correlated words to the same cluster. Thus, to address the problem of coherence of aspect clusters in our scenario, we believe, each cluster of topics must contain the aspects similar to each other. Therefore, the goal of our proposed method is to identify coherent aspects based on the context they share with the other aspects in the cluster.

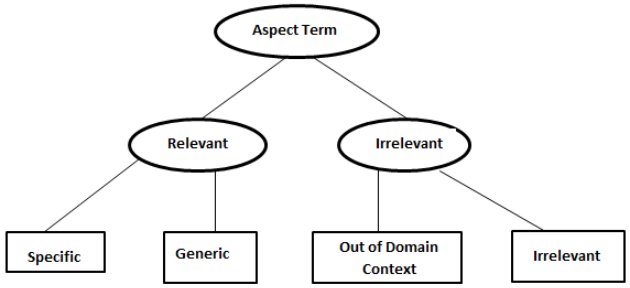

Fig. 1. Example of a topic-based cluster

In order to facilitate the said purpose, we append each topic-based aspect with an automatically generated domain descriptor having context information obtained from the large lexical resource BabelNet. For instance, in Fig. 2, the topic-based aspects 2-11, 13-17, 19 and 20 appear coherent; however topic-based aspects 1, 12 and 18 are not having high coherence with other words in the cluster. We propose BCohen (BabelNet based Coherent topic-based aspect filtering) method which extracts the context domain knowledge from the general-purpose lexical resource BabelNet 1.1.1. This enables the automatic acquisition of related domain list for each topic-based aspect required to construct aspect based domain-descriptors.

Our method further processes these descriptors by computing the dominant domains for each cluster. Our method then identifies the aspects whose descriptors successfully superimpose on the dominant domains as coherent aspects. The main contribution of this paper is to introduce a novel context knowledge based method for automatically building domain-based descriptors for each aspect in the cluster. We show the usefulness of the proposed method by processing the domain-based descriptors for improving the coherence of the topic-based clusters by classifying the aspects as shown in the Fig. 1.

This method of aspect filtration helps in finding relevant aspects and further discriminating them into domain-specific aspects from cross-domain generic aspects.

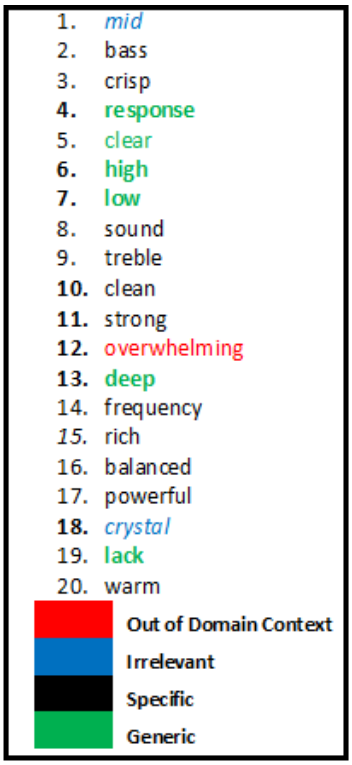

Fig. 2. Example of a topic-based cluster

To the best of our knowledge, this is the first paper that uses domain knowledge for improving coherence of topic clusters and identifying the aspect types.

The remainder of this paper is organized as follows: Section 2 presents related work, section 3 describes the architecture, section 4 gives method and algorithm, section 5 presents experiments and section 6 states conclusion and future work.

2 Related Work

In sentiment analysis, aspect extraction aims at extracting the aspects of an entity in a domain on which opinions have been expressed [11, 19, 20]. With very large volume of data, opinions are often conveyed through latent semantics, which make purely syntactical approaches ineffective. To discover the latent structure, [9] proposed matrix decomposition in their model called Latent Semantic Indexing (LSI).

To substantiate the claims regarding LSI, [17] suggested a generative probabilistic model of the text corpora using Bayesian methods. pLSI proposed by [10], is the probabilistic LSI model where the document expansion is done by adding additional index terms automatically by statistical inference. However, pLSI model does not provide probabilistic model at the level of documents. According to [8], each word is generated from a single topic and different words in a document are generated from different topics.

LDA is a mix membership model that builds on the work of [9] and [10]. Each document is a random mixture of corpus wide topics. Each such latent topic is characterized by distribution over words. [14] clustered the aspects to determine the grouping of semantically related aspects. Therefore, we chose LDA based method. It performs soft clustering of terms in its topic modeling process [3].

[1] proposed a method for generating coherent topics using semi-supervised approach which uses LDA for topic discovery and then considers words having similar probability to belong to the same topic and rejects words that have different probabilities across topics. [1] and [6] incorporate prior knowledge using must-links and cannot-links. These topic models are exploited for aspect extraction because they generate the topics and also group them at the same time [6]. However, these models suffer from the problem of multiple senses as the must-links and cannot-links result in transitive relations. LDA with Multi-Domain Knowledge (MDK-LDA) [7] handles multiple senses by choosing a correct sense represented by an s-set(semantic set), a set of words sharing the same semantic meaning.

The s-set for each aspect in the topic is constructed using WordNet. It works on the idea that, given a set S of s-sets from multiple aspects, S is used to produce coherent topics. To deal with multiple senses, a latent variable s is added to LDA to model s-sets. Further to it, [6] showed that lexical knowledge can be used for predictive coherence improvements of topic-based aspects. [6] used WordNet to derive synonyms and antonyms to improve coherence.

However, this approach did not provide complete context specific information due to the coverage issue of WordNet. In [5] AKL, uses LDA to construct the common topic-base from data across multiple domains. Then, it clusters the topic-based aspects using k-mediod. We construct clusters of topic-based aspects using AKL. We then focus on the use of publicly available lexical resource BabelNet 1.1.1. We propose an approach for improving topic coherence based on the hypothesis which states that topic-based aspects with similar context domains tend to be semantically similar and hence more coherent.

BabelNet is a very large, wide coverage multilingual semantic network built from WordNet and Wikipedia [16]. It supports 3 million concepts. BableNet generates knowledge graphs by combining senses by intersecting relationships from Wikipedia and senses in WordNet. It provides mappings even when the intersection between WordNet and Wikipedia disambiguation contexts is empty [16].

3 Architecture

As a standard approach, an aspect based topic model takes as an input a set of documents and generates sets of topic-based aspects. These topic-based aspects help in expressing hidden thematic structure underlying a dataset. A popular topic model LDA [3] represents a document into a generative probabilistic model. It takes T as the number of topics and a set of Dirichlet hyperparameters. These latent distributions are inferred using Expectation-Maximization or Gibbs sampling, among other approaches. In this section, we describe our architecture.

Our proposed architecture is shown in Fig. 3. We used LDA based topic model with the parameter of number of topics Z set. We thus obtain Z topics. Using these topics, we generated topic-based aspects across multiple domains thereby accumulated in a common topic base. AKL introduced by [5] is a semi-supervised method which performs latent inference using Gibbs sampling. We obtain cluster of topic-based aspects using AKL method. Each cluster thus obtained consists of top k aspects.

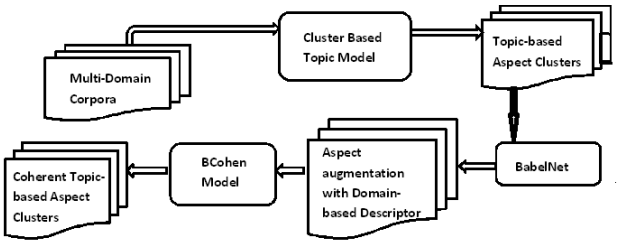

Fig. 3. BCohen Architecture

The key step in the architecture is the approach used to improve coherence of thus obtained clustered topic-based aspects. For this we leverage context domain information using BabelNet 1.1.1. Each aspect in the cluster is sent as an input to BabelNet. For each such input BabelNet returns a list of all the domains to which the aspect likely belongs. We structure this list and call it as a domain-descriptor. Thus, each topic-based aspect in the cluster is augmented with domain-based descriptor. Such aspects are further processed by BCohen method presented in section 4 to improve topic-based cluster coherence by filtering the relevance of aspects in the cluster into various categories namely specific, generic, out of domain context and irrelevant.

4 Method and Algorithm

Every topic-based aspect augmented by domain-based descriptor could be expressed as shown in Equation 1:

$$\{< Top\_ID >< Description >: [< d_1 >, < d_2 >, ..., < d_n >]\}. \quad (1)$$

The elements of the expression are specified as follows:

1. Top_ID is the key for indexing the topics retrieved from the topic base.
2. Description is the identifier specification of the aspect i.e. aspect name.
3. The domain list of the description i.e. d_n domains, the aspect assumes by its context domain knowledge. BCohen refers to this as domain-based descriptor.

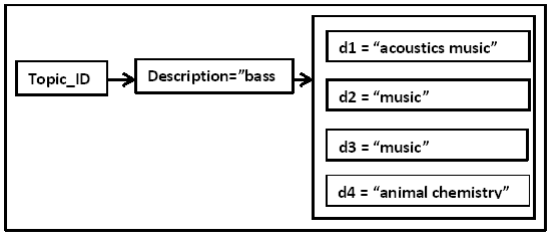

Fig. 4. An instance of topic-based aspect augmented with a domain-based descriptor

An instance of a topic-based aspect in a cluster augmented with a domain-based descriptor is illustrated in Fig. 4. The domains associated with an aspect, obtained from BabelNet, may repeat like "music"in Fig. 4 indicating higher likelihood of influence of that domain on the aspect context.

Clusters of such domain-based descriptor augmented topic-based aspects are processed by BCohen approach which is concretely presented in Table 1 and is described here.

A. Clustering stage:

Step 1.1 **Base topic knowledge generation using LDA.**
To generate the initial topic base LDA is run on the corpus. Generate each word in the document by:

i. Picking the topic z using the multinomial distribution generated from the dirichlet hypergenerator.

ii. Using the topic and the word-topic distribution, word is generated.

Step 1.2 **Generate cluster of topics using AKL model.**
Connect topics and terms using latent cluster c of AKL. Generate each word in a cluster of a document by:

i. For a topic z, using the mixture proportion from the dirichlet hypergenerator, each cluster c is generated.

ii. Using the topic z and the cluster c, the word w is generated.

Step 1.3 **Combine topic clusters from multiple domains to generate a common aspect topic base.**
Each domain now is expressed as collection of clusters of semantically related words. The purpose of clustering was to group semantically related topics. K-medoid method of clustering was effectively used for this purpose in [5]. We propose to use clusters generated by [5] for further effective knowledge discovery for aspect coherence improvement. Two issues raised by AKL are :

i. Aspect discrimination: Need for the discovery of generic aspects shared across the domains and specific aspects tied to a certain domain.

ii. Multiple senses: Ambiguity due to multiple senses associated with similar terms occurring in clusters of different domains.

We propose to intervene at this point as we believe that instead of using the statistical means to derive coherence we could exploit the semantic relations. The idea is that the words occurring in the same context tend to have similar meanings. Thus, every topical aspect is automatically appended by the descriptor specifying domain list based on the context. For this purpose we propose to use domain lexical semantic knowledge from BabelNet 1.1.1.

B. Exploiting topic-based aspect context knowledge for coherence improvement :

Step 2.1 **Express every topical aspect in the cluster by constructing the domain descriptor using BabelNet 1.1.1** For each topical aspect from each cluster, sense tagging based on the context is to be done. For this purpose we use BabelNet 1.1.1. We find the connecting paths on the knowledge graph generated by BabelNet and merge identified paths. We tap all generated likely domains associated with each aspect. Each topical aspect is thus an instance expressed by augmenting

its descriptor comprising of associated domains.

Step 2.2 **Select dominant domains of the cluster by computing the most frequently occurring top n domains.**
Each cluster is a collection of topical aspects which are supposedly semantically similar. In order to conform to the coherence of the aspects we exploit the domain descriptors of each aspect. As a result, the domain matrix for each cluster is constructed. Then, the most frequently occurring top domains are identified as dominant domains. The dominant domain would indicate the semantic inclination of the cluster towards the certain context. The dominant domains derive the common context which semantically relate the words in a cluster.

Step 2.3 **Compute the superimposition of each topical aspect domain descriptor on the dominant domain vector.**
A topical aspect with the domain knowledge from its descriptor is superimposed on the dominant domain vector obtained in step 2.2. The superimposition reflects the semantic inclination of topical aspects which contribute to their coherence as semantically similar aspects are coherent aspects.

Step 2.4 **Eliminate the aspects not supporting the dominant domains as incoherent aspects.**
The topical aspects having zero superimposition are declared as incoherent aspects.

Step 2.5 **Discriminate the coherent topical aspects as generic and specific aspects.**
Topical aspects inherently are inclined either towards the generic or specific domain. Therefore we discriminate the dominant domains into two types generic and specific.

We have presented our complete method in Algorithm 1 and Algorithm 2.

Table 1. Specification of the BCohen approach

Stage 1 : Clustering stage
Input : Multi domain big collection of opinion data D
Output : Clustered topic aspects
Step 1.1 : Topic generation using LDA
$T_B \leftarrow$ Run LDA on each domain $D_i \in D$
Step 1.2 : Generate cluster of topics by using AKL
$T_C \leftarrow$ Connect topics and terms using AKL
Step 1.3 : Combine topic clusters from multiple domains to generate a common aspect topic base.

Stage 2 : Exploiting topic-based aspect context knowledge for coherence improvement
Input : Clusters of topic-based aspects
Output : Coherent Clusters
Step 2.1 : Augment every topic-based aspect in the cluster with the domain-based descriptor list using BabelNet 1.1.1.
Step 2.2 : Construct the dominant domain vector per cluster by computing the most frequently occurring top domains.
Step 2.3 : Compute the superimposition of each domain-based descriptor of the topic-based aspect, on the dominant domain vector.
Step 2.4 : Eliminate the aspects not supporting the dominant domains and declare them as incoherent aspects.
Step 2.5 : Discriminate the domain-specific coherent topical aspects from cross-domain generic topical aspects.

We now describe the methods given in Algorithm 2 which is specific to our proposed BCohen approach.

findDomDomains(C_i)
For each cluster C_i in the product domain, context matrix CM_{ql} is constructed using domain-based descriptors of topic-based aspects. The context matrix CM_{ql} consists of V_i rows X V_j columns. The dominant domain of the cluster is computed as given in the Equation 2:

$$dom_C = max\{V_{ij}\}, \tag{2}$$

where $1 \leq i \leq \mid V_i \mid, 1 \leq j \leq \mid V_j \mid$.

Input : Opinion Domain Data D
Output : Clusters of topic-based aspects C
foreach *opinion domain data* $D_i \in D$ **do**
  $A_i \leftarrow$ AKL(D_i) ;
  $A \leftarrow A \cup A_i$;
end
$C \leftarrow$ Clustering(A);
foreach *cluster* $C_i \in C$ **do**
  $CC_i \leftarrow$ BCohen(C_i);
end

Algorithm 1: Constructing Topic-based Aspect Clusters

We identify the dominant domain associated with V_{ij} as, domDomain$_C$ = ij^{th} entry in the CM_{ql}. Our observation across CM_{ql} of the product clusters noticed that the context domain "factotum" is the dominant domain across clusters. Since "factotum" is the generic context domain name assigned by BabelNet, we conclude that the topical aspects having maximum superimposition on "factotum" domain context will be assigned as generic topical aspects. The next frequently occurring context domain in CM_{ql} is the second dominant domain and so on. Our observations across the clusters through the product concluded that frequently occurring domains are specific to the cluster context and they do indicate semantic relation across the topic-based aspects of that cluster.

superimpose(w_{ql}, C_i)
Within each cluster, each topic-based aspect has its domain-based descriptor w_{ql}. The context domains in w_{ql} are indexed as j $\in$ 1,..,m. We compute the superimposition of w_{ql} on the dominant domain vector D_v based on the frequency of matches as in Equation 3. The match is successful if current context domain in w_{ql} descriptor matches the current dominant domain in D_v:

$$count(w_{ql}, j) = \sum_{k=1}^{|m|} f(w_{ql_k}). \quad (3)$$

The order of superimposition is considered to be most specific to least specific.

5 Experimental Results

In this section we describes working of BCohen model and related results.

5.1 Example

We first present description of a sample instance that is generated by BCohen model in Fig. 1 and the classification of "irrelevant "and "relevant "aspects is shown using color codes. The interpretation of color codes is described as follows:

1. The topic-based aspect shown in red color and bold font: 12 is out of domain context with null context domain-descriptor from BabelNet. BCohen uses this information to imply that such aspect is "irrelevant and out of domain context "and therefore contributes to incoherence of the cluster.
2. The topic-based aspects shown in black color: 2, 3, 8-11, 14-17 and 20 are classified as "relevant and specific "aspects, obtained by superimpose(w_{ql},C_i) function of BCohen. The domain-based descriptors of such aspects find most superimposition on dominant domain vector. Such topic-based aspects contribute towards enhancing the coherence of the topic-based clusters.
3. The topic-based aspects shown in blue and italic font: 1 and 18 are identified as completely "irrelevant "aspects, by the superimpose(w_{ql},C_i) function of BCohen. The domain-based descriptors find least superimposition which tends to zero, on dominant domain vector due to which these aspects contribute to the incoherence of the cluster
4. The topic-based aspects shown in green and bold font: 4-7, 13 and 19 are identified as "relevant and generic "aspects, by the superimpose(w_{ql},C_i) function of BCohen. The domain-based descriptors find "factotum "as a dominant domain, due to which these aspects are not tied to a particular domain and are found occurring across domains.

Input : Clusters of topic-based aspects C from Opinion Domain Data D
Output : Coherent Clusters

```
foreach cluster C_i ∈ C do                          // Build descriptor for each topic-based aspect of cluster
    foreach word w ∈ C_i do
        w_ql ← Babel(w)                             // Construct qualifier list using BabelNet 1.1.1
        w ← w + w_ql                                // Update topic-based aspect
    end
end
foreach cluster C_i ∈ C do                          // Compute dominant domain vector
    D_v ← findDomDomains(C_v)
end
foreach cluster C_i ∈ C do                          // Filter relevant topic-based aspects
    foreach word w ∈ C_i do
        R_v ← superimpose(w_ql , C_i )
    end
end
return R_v
```

Algorithm 2: BCohen algorithm for exploiting topic-based aspect domain-descriptors for coherence improvement

Table 2 shows two sample clusters corresponding to thematic components from Headphone and GPS domains and the classification of aspects done by BCohen is shown using color codes. Each column indicates a topic cluster in the respective domain. The total number of topical aspects per cluster, as stated before, is 20. The colors indicate the type of topical aspect and its respective contribution to cluster coherence as stated in the BCohen instance description presented before.

From Table 2 we can see that BCohen model is able to discover domain specific aspects in black and domain generic aspects in green thereby filtering irrelevant aspects resulting in generation of improved cluster coherence.

5.2 Experimental Evaluation

This section compares BCohen implementation for coherent topic modeling and its impact on topic coherence.

5.2.1 Experimental Settings

For our experiments we used review corpora from amazon.com pertaining to 36 product domains [5]. These datasets contain approximately 1000 reviews from each domain. We use these for generating the topic-based aspects and further for generation of topic-based aspect clusters. In our experiments, we compare BCohen with LDA and AKL, which are been used as baselines. LDA is a standard unsupervised topic model.

For each sample across all models we perform posterior estimation with 250 burn-in iterations for total 2000 iterations. The parameter setting is done with α =1, β =0.1 and γ with number of words in the cluster. We set Z, the number of topics representing clusters to be 14. We set K, the number of topic-based aspects per cluster as 20. We implemented the AKL topic model in Java. Our implementation uses k-mediod clustering as described in the original paper. We used BabelNet for leveraging context domains required for constructing the domain-based descriptors for each topic-based aspect.

Table 2. Sample topical clusters generated from BCohen and AKL

Headphone				GPS			
AKL	BCohen	AKL	BCohen	AKL	BCohen	AKL	BCohen
music	music	ear	ear	port	port	price	**price**
movie	movie	size	size	usb	**usb**	worth	**worth**
file	*file*	bud	*bud*	cable	cable	purchase	purchase
computer	computer	cord	cord	adapter	adapter	money	**money**
player	player	piece	piece	computer	computer	review	review
song	song	wire	wire	connection	connection	device	device
type	**type**	part	part	car	car	model	model
audio	audio	short	**short**	cord	cord	user	user
rock	rock	side	**side**	print	print	magellan	*magellan*
video	video	pad	pad	printer	printer	research	research
portable	portable	canal	**canal**	server	server	range	range
audiophile	**audiophile**	hook	hook	power	**power**	nuvi	**nuvi**
phone	phone	uncomfortable	**uncomfortable**	laptop	laptop	friendly	**friendly**
classical	**classical**	plastic	*plastic*	charger	charger	mine	*mine*
metal	*metal*	cup	cup	lighter	**lighter**	experience	experience
cd	cd	rubber	**rubber**	cigarette	*cigarette*	negative	negative
kind	**kind**	left	**left**	receiver	receiver	glad	**glad**
bass	bass	case	case	serial	**serial**	star	**star**

5.2.2 Measuring Topic Coherence

The topical aspects that co-occur frequently are close to each other within a semantic space and are likely to contribute to higher levels of coherence [18]. Testing was done based on variable number of aspects per topic cluster per domain and the effect of interpretability was recorded based on topic coherence measure. Topic Coherence metric [15] or UMass metric [22] rely upon word co-occurrences and is proposed as a good measure for assessing topic quality. Higher UMass score indicates higher topic coherence indicating better quality.

Fig. 5 shows average UMass score computed for the models given different number of topics and varying aspects per topic. From Fig. 5 we observe that BCohen performs better than LDA and AKL with higher coherence obtained with increasing topics. This shows that BCohen results in more interpretable topics and improvement in coherence is attributed to identification of “irrelevant ”aspects.

Also, our observations noted that initial drop of coherence in BCohen as compared to AKL is due to certain aspects which are not identified as relevant by BCohens induced domain knowledge. For example in GPS domain the occurrences of words like “tomtom ”.

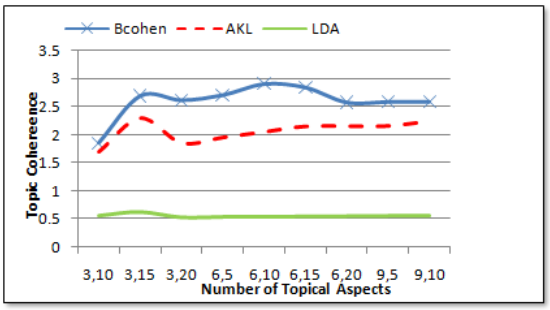

Fig. 5. Topic Coherence

5.2.3 Measuring Generality

We followed the instructions in [18] to evaluate generic aspects identified by BCohen, as such aspects have “factotum ”as a dominant domain by superimposition of BCohen and therefore are shared across topic clusters. Therefore, we are interested in computing the overlap. For this

purpose [18] suggested mean pairwise Jaccard similarity. This metric finds the topical aspects having high occurrence frequency across topic clusters as generic aspects. We observed that Jaccard score increases with the increase in the number of aspects.

Table 3. Mean Pairwise Jaccard Similarity

Number of Topical Aspects	Mean Pairwise Jaccard Score
4	0.032
6	0.0796
8	0.144
10	0.188
12	0.2386
14	0.2446
16	0.2746
18	0.3314
20	0.382

5.2.4 Measuring Cluster Quality

BCohen has the ability of classification of topic aspects at the intra-cluster level which is based on external labels. The determination of quality of classification in a topic cluster is done by entropy measure. As per the information theory, minimal possible degree of disorder is indicated by minimum entropy [21].

Entropy measure evaluates average information content computing degree of distribution of various classes within a cluster [21]. For a particular cluster C_x of size n_x, the entropy of the cluster is given in the Equation 4:

$$E(C_x) = -\sum_{i=1}^{n_c}(\frac{n_x^i}{n_x} log \frac{n_x^i}{n_x}). \quad (4)$$

In Equation 4, n_c are the number of classes and n_i^x refer to the number of instances of the i^{th} class that belong to the x^{th} cluster.

With BCohen, the classes identified at the intra-cluster level are shown using color code in Fig. 6. Entropy computation in BCohen method is based on probability distributions of topic words in a cluster and the classes are known apriori .

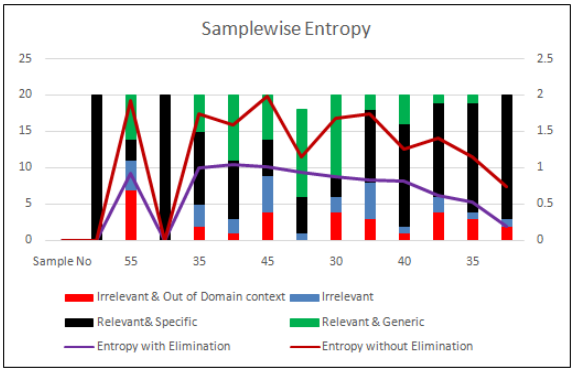

Fig. 6. Example of a topic-based cluster

The classes indicated by color codes in Fig. 6, show distribution of each class. As observed in Fig. 6, the consequence of such a principle is that when the probability distributions generate entropy values greater than 1 as the number of classes is more than 2. Fig. 6, shows that the entropy with elimination of irrelevant topic aspects is lower than the entropy considering all aspects without elimination. Therefore, the elimination of topical aspects which are irrelevant and irrelevant due to out of domain context from topic clusters, offer higher quality clusters due to lower disorder. Also, Fig. 6 shows that the clusters where all aspects belong to the same class have entropy tendind to 0 as such clusters are homogenous.

6 Conclusion and Future Work

We implemented a topic-based aspect extraction and clustering model based on the past work. Such topic-based aspect clusters were obtained from multi-domain opinion text data collection. We observed that not all topic-aspects in thus created topic-based clusters comply with the common context and hence, the need for improvement of the coherence of clusters was addressed. We introduced our lexical domain knowledge based approach which uses BabelNet - lexicalized semantic network resource, to leverage domain context knowledge associated with topic-based aspects.

We introduced BCohen approach which processes the topic-aspects augmented with domain

context knowledge to filter the incoherent aspects thereby improving the cluster coherence. On experimentation with amazon multi-domain review datasets, we observed that BCohen results in the improved topic coherence supported by higher UMass score and lower entropy in cluster quality results in better performance. As compared to our baseline AKL, BCohen scales linearly with the increase in the number of the context domains and it scales in cubic range with the increase in data size. The performance of BCohen on real datasets suggest that our approach could be an interesting tool in practice, as it can be used to filter coherent aspects automatically based on the context domain knowledge. Our paper points to the fact that our approach can be extended in future to multilingual settings for improving coherence of multilingual clusters of topic-based aspects as BabelNet is a multilingual lexical resource.

References

1. **Andrzejewski, D., Zhu, X., & Craven, M. (2009).** Incorporating domain knowledge into topic modeling via Dirichlet forest priors. *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM, pp. 25–32.
2. **Bing, L., Jiang, S., Lam, W., Zhang, Y., & Jameel, S. (2015).** Adaptive concept resolution for document representation and its applications in text mining. *Knowledge-Based Systems*, Vol. 74, pp. 1–13.
3. **Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003).** Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022.
4. **Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009).** Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, pp. 288–296.
5. **Chen, Z., Mukherjee, A., & Liu, B. (2014).** Aspect extraction with automated prior knowledge learning. *Proceedings of ACL*, pp. 347–358.
6. **Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R. (2013).** Discovering coherent topics using general knowledge. *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, ACM, pp. 209–218.
7. **Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R. (2013).** Leveraging multi-domain prior knowledge in topic models. *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, AAAI Press, pp. 2071–2077.
8. **De Finetti, B. (1990).** Theory of probability, volume I.
9. **Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990).** Indexing by latent semantic analysis. *JAsIs*, Vol. 41, No. 6, pp. 391–407.
10. **Hofmann, T. (1999).** Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 50–57.
11. **Hu, M. & Liu, B. (2004).** Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 168–177.
12. **Kumar, A., Kumar, N., Hussain, M., Chaudhury, S., & Agarwal, S. (2014).** Semantic clustering-based cross-domain recommendation. *Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on*, IEEE, pp. 137–141.
13. **Liu, B. (2012).** Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, Vol. 5, No. 1, pp. 1–167.
14. **Mauge, K., Rohanimanesh, K., & Ruvini, J.-D. (2012).** Structuring e-commerce inventory. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, Association for Computational Linguistics, pp. 805–814.
15. **Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011).** Optimizing semantic coherence in topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 262–272.
16. **Muthén, B. (2003).** Statistical and substantive checking in growth mixture modeling: comment on Bauer and Curran.
17. **Navigli, R. & Ponzetto, S. P. (2012).** Multilingual WSD with just a few lines of code: the BabelNet API. *Proceedings of the ACL 2012 System Demonstrations*, Association for Computational Linguistics, pp. 67–72.

18. **O'Callaghan, D., Greene, D., Carthy, J., & Cunningham, P. (2015).** An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, Vol. 42, No. 13, pp. 5645–5657.
19. **Papadimitriou, C. H. & Steiglitz, K. (1998).** *Combinatorial optimization: algorithms and complexity*. Courier Corporation.
20. **Popescu, A.-M. & Etzioni, O. (2007).** Extracting product features and opinions from reviews. In *Natural language processing and text mining*. Springer, pp. 9–28.
21. **Shannon, C. E. (1948).** A mathematical theory of communication. *Bell system technical journal*, Vol. 27, No. 3, pp. 379–423.
22. **Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012).** Exploring topic coherence over many models and many topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Association for Computational Linguistics, pp. 952–961.
23. **Titov, I. & McDonald, R. (2008).** A joint model of text and aspect ratings for sentiment summarization. *Proceedings of ACL-08: HLT*, pp. 308–316.

Article received on 08/06/2016; accepted on 02/04/2018.
Corresponding author is Kavita Asnani.