Robust Multi-view Topic Modeling by Incorporating Detecting Anomalies

Guoxi Zhang[1], Tomoharu Iwata[2], and Hisashi Kashima[3]

[1] Graduate School of Informatics, Kyoto Universityb
guoxi@ml.ist.i.kyoto-u.ac.jp
[2] NTT Communication Science Laboratories
iwata.tomoharu@lab.ntt.co.jp
[3] Graduate School of Informatics, Kyoto University
kashima@i.kyoto-u.ac.jp

Abstract. Multi-view text data consist of texts from different sources. For instance, multilingual Wikipedia corpora contain articles in different languages which are created by different group of users. Because multi-view text data are often created in distributed fashion, information from different sources may not be consistent. Such inconsistency introduce noise to analysis of such kind of data. In this paper, we propose a probabilistic topic model for multi-view data, which is robust against noise. The proposed model can also be used for detecting anomalies. In our experiments on Wikipedia data sets, the proposed model is more robust than existing multi-view topic models in terms of held-out perplexity.

1 Introduction

Multi-view text data consist of texts from different information sources. A view of an instance refers to a part that is from some information source. For example, in a English-Japanese bilingual corpora, an document has two views: English article and Japanese article. Multi-view text data are considered as comparable if views of a document are description of the same target. Multi-view topic modeling is the task of extracting aligned topics from comparable multi-view text data, which are tuples of semantically similar topics of different views. Aligned topics facilitate construction of bilingual lexicon of semantically related words, which can be useful in cross-lingual information retrieval [16]. Aligned topics can also be used to transfer knowledge from one language to another in cross-lingual document classification [5]. Moreover, on data consist of texts and social annotation, complementary information in tags can be utilized to improve performance of clustering tasks [14] using aligned topics.

In a multi-view topic models, a view of a document is modeled as a mixture of topics, which are Categorical distributions over words. The mixture weights, which are often called topic proportions, can be considered as low-dimensional representation of documents. It is often assumed in existing multi-view topic models that different views of the same document are semantically consistent. Under this assumption, topic proportions are shared across all views of a document [11]. However, for multilingual corpora that are managed in distributed fashion, this assumption does not necessarily hold. For example, since articles in different Wikipedia languages are usually managed

by different communities, they often differ in details. Figure 6 shows an example for this. This bilingual document contains Japanese article and Finnish article about Orne, a province in France. Compared to Japanese article, Finnish article contains more information about history of Orne, so it should have larger weights for topics that related to history.

Documents that have inconsistent weights can be regarded as multi-view anomalies [8]. Although inconsistency in content should incur difference in topic proportions, existing models are not capable of depicting it while learning low-dimensional representation of documents. In this paper we propose a multi-view topic model which models data and detects anomalous instances simultaneously. Appropriate number of topic proportions variable are inferred for anomalous instances, to model the inconsistent views. As a result, the proposed model is more robust to multi-view anomalies, and also applicable for the multi-view anomaly detection task. The proposed model is beneficial in at least two applications. In large enterprise with global business, managing information consistency in multilingual documents is an important but expensive task [6]. Cost of management can be reduced if anomalous documents are detected automatically. In cross-cultural analysis [9], documents with inconsistent views are used to analyze cultural difference. We can reduce cost of obtaining samples by using the proposed model to identify anomalous documents from large datasets automatically.

In the proposed model, documents that contain inconsistent views are regarded as anomalies, and such views have distinct topic proportions. Views of a non-anomalous documents share the same topic proportions variable. We use Dirichlet process as the prior for topic proportions variables to infer the appropriate number of topic proportions variable for each document. Based on collapsed Gibbs sampling, we derive efficient inference procedures for the proposed model. To our knowledge, this is the first model that addresses the problem of multi-view anomaly detection in the literature of topic modeling. Performance of the proposed model is examined on ten bilingual Wikipedia corpora. It is demonstrated that the proposed model is more robust than existing multi-view topic models, in terms of held-out perplexity. In addition, compared to existing multi-view anomaly detection methods the proposed model is more efficient and has higher anomaly detection performance on multi-view text data.

The rest of this paper is organized as the following. Section 2 includes related work on topic modeling and multi-view anomaly detection. The proposed model and its inference method are presented in section 3. Section 4 contains evaluation of models' generalization ability in terms of held-out perplexity on Wikipedia corpora. Section 5 contains evaluation of multi-view anomaly detection. In Section 6, examples of aligned topics and multi-view anomalies in a Wikipedia corpus are presented. We conclude this paper in section 7.

2 Related Work

Topic models, such as Latent Dirichlet Allocation (LDA) [4], are analysis tool for discrete data. Polylingual Topic Model (PLTM) [11] is an extension of LDA to comparable multi-view setting, and is demonstrated to be useful in various applications, such as cross-cultural understanding, cross-lingual semantic similarity calculation and cross-

lingual document classification [16]. Based on the fact that views of a document are information about the same target from different perspective, topics of different views are aligned by sharing topic proportions variable among all views of a document in PLTM. While information in different views are utilized jointly, this model assumption is not valid for data that contain multi-view anomalies. Correspondence LDA [3] and symmetric correspondence LDA [7] are another kinds of multi-view topic models, which extract direct correspondence between topics of different views. However, in these models distinct topic proportions variables are inferred for views of the same document, in regardless of view consistency. Hence they are not applicable in detecting multi-view anomalies and obtaining low-dimensional representation of multi-view documents. Moreover, in existing models topics are to be aligned without considering view consistency, so on noisy data that contains a lot of multi-view anomalies their performance may degenerate.

Various methods can be applied to the task of multi-view anomaly detection. In probabilistic canonical correlation analysis (PCCA) [2] a shared latent vector among all views and its projection matrices for each view are estimated. The reconstruction error is considered as anomaly score, based on the idea that high reconstruction error indicates views are inconsistent. In [10] the authors propose a robust version of PCCA by detecting multi-view anomalies during estimating parameters. Nevertheless, this model assumes Gaussian error so it may not be suitable for textual data. Moreover, textual data have high-dimensional features, which leads to efficiency issue when applying that method.

3 Proposed Model

3.1 Generative Process

Suppose there are D documents, and each of them contains L views. In the proposed model, views of a document are grouped into clusters. The proposed model assumes that each document can have a countably infinite number of clusters. A topic proportion vector θ_{dy} is generated for each cluster y in document d, and it is then used to generate words in each view of y. As a result, views in the same cluster share the same topic proportions vector, and views belong to different cluster have distinct topic proportions vectors. Consequently, multi-view anomalies are identified by the number of clusters they have. A document is a normal document if it has only one cluster, and is an anomaly if it has more than one cluster.

Specifically, we use Stick-breaking process [15] to generate clusters and cluster assignments of views. The probability that a view belongs to some cluster is related to the proportions of its words' topic assignments. In anomalous documents, such proportions are different in different views, causing its views to be assigned to different clusters. Meanwhile, in normal documents such proportions of views are similar, so views are assigned to the same cluster.

The generative process of the proposed model is described as the following, and the graphical model representation is shown in Figure 1.

For each $\ell = 1, 2, \ldots, L$ and $k = 1, 2, \ldots, K$, generate a topic $\phi_{\ell k} \in R^{V_\ell}$ with a symmetric prior $\beta_\ell e \in R^K$, where $\beta_\ell \in R$ and $e \in R^K$ is all-ones vector. V_ℓ is the

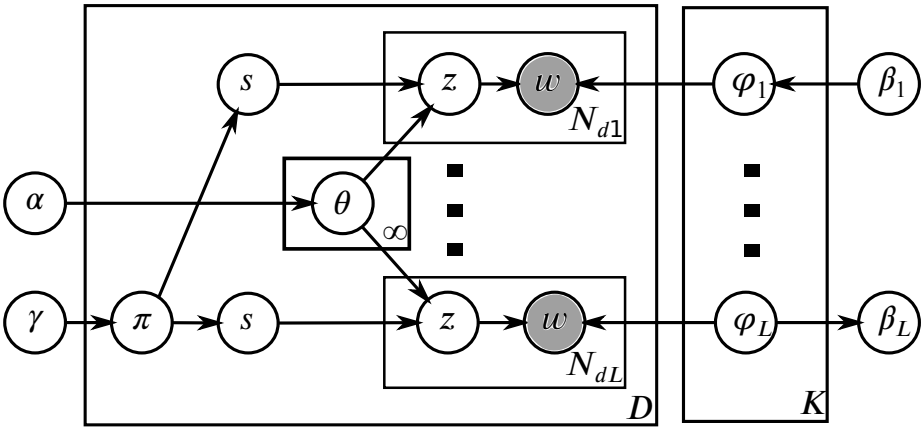

Fig. 1: Graphical model representation of the proposed model.

number of unique words in view ℓ.

$$\phi_{\ell k} \sim \text{Dirichlet}(\beta_\ell). \tag{1}$$

For each document d, generate mixture weights π_d by the stick-breaking process with concentration parameter γ, which generates mixture wights of the Dirichlet process.

$$\pi_d \sim \text{Stick}(\gamma). \tag{2}$$

For each view ℓ of the document d, generate cluster assignments $s_{d\ell}$ from π_d:

$$s_{d\ell} \sim \text{Category}(\pi_d). \tag{3}$$

Then generate topic proportions θ_{dy} for cluster y of d using asymmetric prior $\alpha \in R^K$:

$$\theta_{dy} \sim \text{Dirichlet}(\alpha). \tag{4}$$

Finally, generate topic assignment $z_{d\ell n}$ of n^{th} word in view ℓ of d, and the corresponding word $w_{d\ell n}$, for $n = 1, \ldots, N_{d\ell}$, where $N_{d\ell}$ is the number of words in view ℓ of document d.

$$z_{d\ell n} \sim \text{Dirichlet}(\theta_{ds_{d\ell}}), \tag{5}$$

$$w_{d\ell n} \sim \text{Category}(\phi_{\ell z_{d\ell n}}). \tag{6}$$

3.2 Inference

Collapsed Gibbs Sampling In the following inference procedure, θ, π and ϕ are marginalized out by Dirichlet-multinomial conjugacy. Denote Z as the topic assignments of all words. Denote S_d as the cluster assignments of all views in document d and S as the identity of cluster assignments in all documents. In order to simplify

expression, denote subscript $d\ell n$ as J, and use $\setminus J$ to refer to the remaining after removing $z_{d\ell n}$. Similarly, use $\setminus d\ell$ to refer to the remaining of a cluster in d after view ℓ is removed. For example, $y \setminus d\ell$ refers to the rest of cluster y after removing view ℓ. If view ℓ is not in y, then y is not modified.

Given S and $Z_{\setminus J}$, Eq.7 is used to sample a new value for $z_{d\ell n}$. Denote the number of occurrence that word t in view ℓ is assigned to topic k as $N_{\ell kt}$. Use $N_{d\ell k}$ and N_{dyk} to refer to number of words in ℓ and in y that are assigned to topic k in document d. Denote number of words in view ℓ^{th} that are assigned to topic k as $N_{\ell k}$.

$$P(z_{d\ell n} = k \mid Z_{\setminus J}, S) \propto (N_{ds_{d\ell}k\setminus J} + \alpha_k) \frac{N_{\ell k w_{d\ell n}\setminus J} + \beta_\ell}{N_{\ell k\setminus J} + \beta_\ell V_\ell}. \tag{7}$$

For each document d, given Z and $S_{d\setminus d\ell}$, Eq.8 is used for sampling a new value for $s_{d\ell}$. ℓ, a view of document d, could be assigned to an existing cluster y or a new cluster $\tilde{y}$. Denote number of words y contains as N_{dy} and number of words in y that are assigned to topic k as N_{dyk}. Denote number of views in cluster y of document d as L_{dy}. $\bar{\alpha} = \sum_{k=1}^{K} \alpha_k$. $\Gamma(\cdot)$ refers to the gamma function.

$$\begin{aligned}
&P(s_{d\ell} = y \mid Z, S_{d\setminus d\ell}) \propto L_{dy\setminus d\ell} \\
&\times \left[\prod_{k:N_{d\ell k}>0} \frac{\Gamma(N_{dyk\setminus d\ell} + N_{d\ell k} + \alpha_k)}{\Gamma(N_{dyk\setminus d\ell} + \alpha_k)} \right] \frac{\Gamma(N_{dy\setminus d\ell} + \bar{\alpha})}{\Gamma(N_{dy} + \bar{\alpha})}, \\
&P(s_{d\ell} = \tilde{y} \mid Z, S_{d\setminus d\ell}) \propto \gamma \\
&\times \left[\prod_{k:N_{d\ell k}>0} \frac{\Gamma(N_{d\ell k} + \alpha_k)}{\Gamma(\alpha_k)} \right] \frac{\Gamma(\bar{\alpha})}{\Gamma(N_{d\ell} + \bar{\alpha})}.
\end{aligned} \tag{8}$$

Hyper-parameter Estimation Hyper parameters α and β smooth word counts in inference. They can be either set to some small values or optimized by placing Gamma priors on them and then using fixed-point iteration method [12]. As demonstrated in [1], the later approach reduce performance difference that is resulted from learning algorithm. Thus we optimize these hyper parameters using the approached introduced in [12], as Eq.9. Y_d denotes the set of clusters in document d. $\Psi(\cdot)$ refers to the digamma function.

$$\begin{aligned}
\alpha_k^{\text{new}} &= \alpha_k \frac{\sum_{d=1}^{D} (\sum_{y \in Y_d} \Psi(N_{dyk} + \alpha_k) - |Y_d| \Psi(\alpha_k))}{\sum_{d=1}^{D} (\sum_{y \in Y_d} \Psi(N_{dy} + \bar{\alpha}) - N_{dy} |Y_d| \Psi(\bar{\alpha}))} \\
\beta_\ell^{\text{new}} &= \beta_\ell \frac{\sum_{k=1}^{K} \sum_{t=1}^{V_\ell} \Psi(N_{\ell kt} + \beta_\ell) - K V_\ell \Psi(\beta_\ell)}{V_\ell \sum_{k=1}^{K} \Psi(N_{\ell k} + V_\ell \beta_\ell) - K V_\ell \Psi(\beta_\ell)}
\end{aligned} \tag{9}$$

Estimation of Θ and Φ After iteratively sampling and updating hyper-parameters, point estimates for Θ and Φ are made:

$$\begin{aligned} \theta_{yk} &= \frac{N_{dyk} + \alpha_k}{N_{ds} + \bar{\alpha}}, \\ \phi_{\ell kt} &= \frac{\beta_\ell + N_{\ell kt}}{N_{\ell k} + V_\ell \beta_\ell}. \end{aligned} \tag{10}$$

Anomaly Score Because view consistency is modeled stochastically using the Dirichlet process, we use the probability that a document has more than one clusters as anomaly score. High value indicates views in a document tend to diverge, so probably it is a multi-view anomaly. As shown Eq.11, such anomaly score is estimated with samples of S generated using the Gibbs sampler Eq. 8. T refers the total number of iterations in model training. $\left|Y_d^{(t)}\right|$ is the number of clusters in document d in iteration t. $\mathrm{I}(\cdot)$ is the indicator function. In experiments we use sufficiently large T to ensure $score_d$ converges.

$$\mathrm{score}_d = \frac{1}{T} \sum_{t=1}^{T} \mathrm{I}(\left|Y_d^{(t)}\right| > 1), \tag{11}$$

4 Held-out Perplexity Evaluation

4.1 Dataset

We collect 34024 articles in Japanese, German, French, Italian, English, and Finnish from Wikipedia. This data is preprocessed by removing general stop words and corpus stop words, which are words with frequency larger than 3402. We also remove words with frequency lower than 100 to reduce the size of vocabulary. After preprocessing, the vocabularies of each language are 12148, 17375, 12813, 16291, 22500, and 7910. From this corpus we select ten bilingual corpora for experiments. They are Japanese - Finnish, Japanese - German, Japanese - French, Japanese - Italian, Japanese - English, English - Germany, English - Finnish, English - Japanese, English - French and English - Italian. We filter out article pairs whose both views are shorter than five words. The numbers of documents in these ten bilingual corpora are 33652, 33668, 33658, 33653, 33854, 33829, 33813, 33854, 33822, and 33814. From each bilingual corpus ten datasets are randomly sampled for experiments, each of them contains 5000 documents.

To quantitively examine models' performance when multi-view anomalies are present, view-swapping is performed to generate multi-view anomalies, as used in [8] and [10]. Specifically, $10\%, 20\%, 30\%, 40\%, 50\%$ of documents in each dataset are randomly selected as anomalies, and their views are swapped . As a result, these datasets contain multi-view anomalies with ratio $10\%, ..., 50\%$. Because data of each view are not modified, these datasets can be used to investigate model's performance against multi-view anomalies.

4.2 Settings

Perplexity of held-out corpus is selected as an evaluation metric. Low perplexity indicates good generalization ability. The proposed model is compared with PLTM and CorrLDA to examine the effect of anomaly detection in multi-view topic modeling.

Perplexity is calculated using Eq. 12. As perplexity of CorrLDA depends on choice of pivot view, we report the average of for different choice of pivot view. The held-out corpus is constructed by randomly selecting 20% of documents and then randomly selecting half of their words in each view. Denote the set of index of documents chosen as D^{test}. Denote the set of words chosen in document d as w_d^{test}. Denote the total number of words chosen as N^{test}.

$$\text{perplexity} = \exp\left(-\frac{\sum_{d\in D^{\text{test}}}\sum_{\ell=1}^{L}\sum_{t\in w_{d\ell}^{\text{test}}}\ln(\sum_{k=1}^{K}\theta_{ds_{d\ell}k}\phi_{\ell kt})}{N^{\text{test}}}\right) \tag{12}$$

In all experiments, initial value of α_k, β and γ are set to 0.05. Gibbs sampling is executed for 1000 iterations. The proposed model is initialized by using single cluster for every documents in the first 256 iterations. After that parameters are learned using procedures described in section 3.2.

4.3 Results

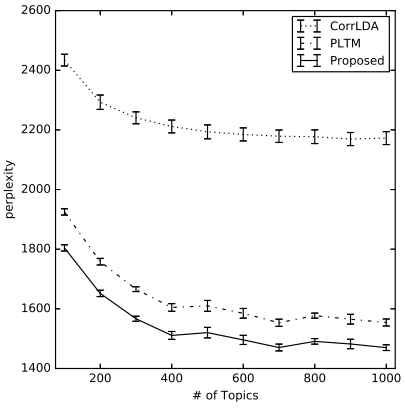

Fig. 2: Average held-out perplexities and their standard errors on Japanese - Finnish dataset with 30% multi-view anomalies

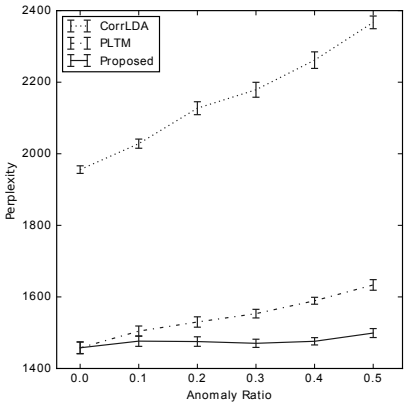

Fig. 3: Average held-out perplexities and their standard error on Japanese - Finnish dataset when anomaly ratio varies. K = 700.

Figure 2 shows the average of held-out perplexities and their standard errors on Japanese - Finnish dataset containing 30% multi-view anomalies. Number of topic K

varies from 100 to 700. With the same K, the proposed model always achieves the lowest perplexity. As perplexities stop decreasing after $K \geq 700$, further increasing number of topics provides no improvement generalization ability. Thus when multi-view anomalies exist, the proposed model outperforms all alternative methods in irrespective of number of topics.

Figure 3 shows average held-out perplexities and their standard errors on the Japanese - Finnish corpus. Anomaly ratio varies from 0 to 0.5. It is shown that as the anomaly ratio increases, perplexities of CorrLDA and PLTM increase significantly. Because view-swapping does not modify content of each view, this performance degeneration could only result from inconsistency among views. Meanwhile, perplexity of the proposed model increases very slowly when the anomaly ratio increases. Note that in Figure 2, the proposed model has the lowest perplexity in regardless of K. We conclude that the proposed model has the best generative ability when multi-view anomalies are present on this bilingual dataset.

Table 1: Average held-out perplexities and their standard errors on 10 bilingual corpora

	English - Germany	English - Finnish	English - French	English - Italian
Proposed	**2828.3±16.8**	**2505.9±11.4**	**2529.3±12.3**	**2664.6±13.9**
PLTM	3036.9±18.1	2602.1±18.7	2664.9±17.93	2841.5±25.3
	English - Japanese	Japanese - Germany	Japanese - Finnish	Japanese - English
Proposed	**2215.3±11.4**	**2130.3± 16.8**	**1470.3±11.5**	**2226.0±13.4**
PLTM	2399.7±18.0	2306.4±13.9	1553.3± 12.0	2368.3±12.5
	Japanese - French	Japanese - Italian		
Proposed	**1728.4±11.2**	**1940.9±15.4**		
PLTM	1866.9±15.2	2130.9±14.2		

Datasets contain 30% multi-view anomalies. K = 700.

Table 1 shows average held-out perplexities and their standard errors on all the ten bilingual corpora with 30% multi-view anomalies for K equals to 700. As shown in Figure 2 and Figure 3, CorrLDA is not suitable for these corpora, so its results are not reported. On all corpora held-out perplexities of the proposed model are significantly lower than those of PLTM. Hence the proposed model's superiority over PLTM on noisy multilingual corpora is language-independent.

5 Multi-view Anomaly Detection

5.1 Settings

Area under ROC curve (AUC) is used as evaluation metric for multi-view anomaly detection. High AUC indicates a method could discriminate anomalous instances from non-anomalous instance well.

The proposed model is compared with robust version of CCA proposed in [10] (RCCA), one-class SVM (OCSVM) and PLTM. RCCA is included in comparison because it also uses Dirichlet process and is reported to be effective on continuos data. OCSVM is a representative method for single-view anomaly detection. It is included into experiments to investigate whether methods for single view anomaly detection are also applicable for detecting multi-view anomalies. In experiments OCSVM implementation in scikit-learn package [13] with radial-basis function kernel is used. It is applied into multilingual setting by using bag-of-word representation and appending one view at the end of another.

We also report results of classification by using PLTM's perplexity of training data as anomaly score. Inasmuch as model assumption of PLTM is not valid on anomalous documents, perplexities of such documents are higher than non-anomalous documents. Because the proposed model reduces to PLTM if cluster assignments of views are fixed to be the same, comparison between the proposed model and this method demonstrates the efficacy of using Dirichlet process to model view consistency.

5.2 Dataset

As RCCA does not scale well on high dimensional textual data, we have to carry out comparison experiments with data of smaller size. On Japanese - Finnish dataset, sizes of vocabulary are reduced 100 by removing low-frequency words. Documents that have view shorter than 50 words are removed. From the remaining ten datasets are sampled, each of them contains 100 documents.

5.3 Results

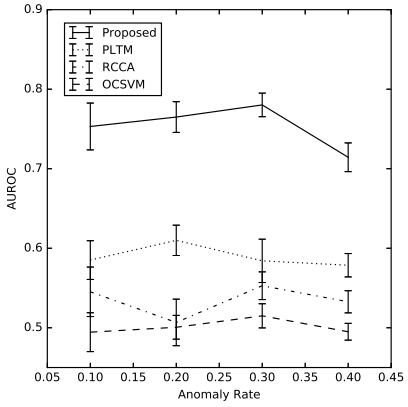

Fig. 4: AUC and their standard errors. Dimension of latent spaces are set to 8 for RCCA. $K = 8$ is used for the proposed model and PLTM

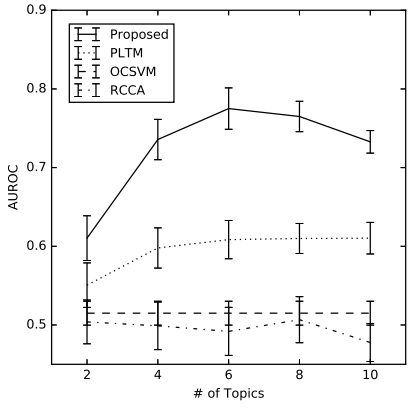

Fig. 5: AUC and their standard errors with anomaly rate equals to 20%.

Figure 4 shows AUC of multi-view anomaly detection when anomaly ratio varies. AUC of RCCA and OCSVM are around 0.5 in all cases, which means they barely discriminate anomalies from non-anomalies. AUC of PLTM is around 0.6, and that of the proposed method is around 0.7. Thus the proposed model outperforms all alternative methods.

Figure 5 shows AUC on dataset containing 20% anomalies with various number of topics K. K correspond to dimension of latent space in RCCA. It is shown that $K = 4$ is enough for the proposed model and PLTM to achieve their best performance, and the proposed method outperform all comparing methods. Meanwhile, AUC of RCCA is around 0.5 for all cases, which means increasing dimension of latent space cannot improve performance of anomaly detection.

6 Examples of Aligned Topics and Anomalies

In previous sections we demonstrate the proposed model's efficacy in modeling multi-view text data with manually created multi-view anomalies and detecting such anomalies. In this section we present topics extracted by the proposed model and example of multi-view anomalies detected from the original data.

fi	final, fantasy, Nintendo, iv, crystal
ja	magic, final, fantasy, character, combat
fi	cooperate, business, production, economics, formula
ja	capital, company, market, analysis, cost
fi	married, spatula, marry, wife, son
ja	marriage, girlfriend, father, mother, daughter
fi	team, score, minutes, world, seconds
ja	team, acting, competition, jump, skate

Table 2: An example of aligned Finnish(fi) and Japanese(ja) topics.

Examples of most probable words of aligned topics extracted from original Japanese - Finnish corpus are presented in Table 2. Relatedness between Japanese topics and Finnish topics are observable. For example, the second topic is about business. With these aligned topics, information of two views can be jointly utilized. For example, the most probable words for fourth Finnish topics are "team", "score", "minutes", "world" and "seconds", which may not be as cohesive as the other topics. It can be better interpreted if the corresponding Japanese topic ("team", "acting", "competition", "jump" "skate") is considered jointly. With the complementary information, one may figure out that words in this topic are about sports competitions.

In addition, an example of multi-view anomaly detected from original Japanese - Finnish corpus is shown in Figure 6. Screenshots are captured in Feb 11th, 2017. These two articles are about Orne, a province in France. While they contain common sections, Finnish and Japanese articles differ significantly in history section. For applications

Fig. 6: Article for Orne in Finnish (left) and its counterpart in Japanese (right).

in which inconsistency is detrimental, we can use the proposed model to detect and process documents like this automatically.

7 Conclusion

Since multi-view text data are often managed in distributed fashion, they may contain multi-view anomalies and pose challenge on topic modeling. In this paper a probabilistic topic model is proposed for multi-view topic modeling, which is capable of modeling joint distribution of views and detecting anomalies simultaneously. In our experiments on ten bilingual Wikipedia corpora, it is demonstrated that the proposed model is more robust than existing multi-view topic models against multi-view anomalies. In addition, from comparison with other multi-view anomaly detection methods it is shown that the proposed model is more effective on textual data. Future work of the proposed model includes applying to multi-modal text data.

References

1. A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 27–34. AUAI Press, 2009.
2. F. R. Bach and M. I. Jordan. A probabilistic interpretation of canonical correlation analysis. *Technical Report 688. Department of Statistics, University of California, Berkeley*, 2005.
3. D. M. Blei and M. I. Jordan. Modeling annotated data. In *Proceedings of the 26th annual International ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134. ACM, 2003.
4. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

5. W. De Smet, J. Tang, and M.-F. Moens. Knowledge transfer across multilingual corpora via latent topics. In *Advances in Knowledge Discovery and Data Mining*, pages 549–560. Springer, 2011.
6. K. Duh, C.-M. A. Yeung, T. Iwata, and M. Nagata. Managing information disparity in multilingual document collections. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(1):1, 2013.
7. K. Fukumasu, K. Eguchi, and E. P. Xing. Symmetric correspondence topic models for multilingual text analysis. In *Advances in Neural Information Processing Systems 25*, pages 1286–1294. Curran Associates, Inc., 2012.
8. J. Gao, W. Fan, D. Turaga, S. Parthasarathy, and J. Han. A spectral framework for detecting inconsistency across multi-source object relationships. In *2011 IEEE 11th International Conference on Data Mining (ICDM)*, pages 1050–1055. IEEE, 2011.
9. N. Hara, P. Shachaf, and K. F. Hew. Cross-cultural analysis of the wikipedia community. *Journal of the American Society for Information Science and Technology*, 61(10):2097–2108, 2010.
10. T. Iwata and M. Yamada. Multi-view anomaly detection via robust probabilistic latent variable models. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1136–1144. Curran Associates, Inc., 2016.
11. D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 880–889. Association for Computational Linguistics, 2009.
12. T. P. Minka. Estimating a dirichlet distribution. https://tminka.github.io/papers/dirichlet/, 2000. Accessed: 2017-01-10.
13. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
14. D. Ramage, P. Heymann, C. D. Manning, and H. Garcia-Molina. Clustering the tagged web. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 54–63. ACM, 2009.
15. J. Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, pages 639–650, 1994.
16. I. Vulić, W. De Smet, J. Tang, and M.-F. Moens. Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management*, 51(1):111–147, 2015.