Relating Romanized Comments to News Articles by Inferring Multi-*glyphic* Topical Correspondence

Goutham Tholpadi and **Mrinal Kanti Das** and **Trapit Bansal** and **Chiranjib Bhattacharyya**
Indian Institute of Science
Bengaluru, INDIA 560004
{gtholpadi,mrinal,trapit,chiru}@csa.iisc.ernet.in

Abstract

Commenting is a popular facility provided by news sites. Analyzing such user-generated content has recently attracted research interest. However, in multilingual societies such as India, analyzing such user-generated content is hard due to several reasons: (1) There are more than 20 official languages but linguistic resources are available mainly for Hindi. It is observed that people frequently use romanized text as it is easy and quick using an English keyboard, resulting in multi-*glyphic* comments, where the texts are in the *same language* but in *different scripts*. Such romanized texts are almost unexplored in machine learning so far. (2) In many cases, comments are made on a specific part of the article rather than the topic of the entire article. Off-the-shelf methods such as correspondence LDA are insufficient to model such relationships between articles and comments. In this paper, we extend the notion of correspondence to model multi-lingual, multi-script, and inter-lingual topics in a unified probabilistic model called the Multi-glyphic Correspondence Topic Model (MCTM). Using several metrics, we verify our approach and show that it improves over the state-of-the-art.

1 Introduction

Analyzing comments on news articles can be cast as modeling correspondence between two sets of variables. Supervised methods do not scale due to unavailability of labeled datasets and rapid growth in unlabeled datasets. Unsupervised methods based on topic models are more appropriate, e.g. correspondence has been explored earlier for images–tags (Blei and Jordan 2003), and articles–comments (Das, Bansal, and Bhattacharyya 2014).

The motivation for this work stems from the problem of analyzing comments in multilingual environments such as in India. This is a hard problem due to several reasons:
(1) There are several nuances to the relationship between comments and articles. While many comments are related to the general topic of the article (*general* comments), some comments relate to a specific part of the article (*specific* comments). In many cases, comments talk about things unrelated to the article. Sometimes, the comment may seem close to topic of the article but was actually made with malicious intent to spam. We call such comments *irrelevant* comments.
(2) Several languages are simultaneously used in multilingual communities (e.g. there are more than 20 official languages in India (Wikipedia 2014)) but linguistic tools are available mainly for Hindi. Users comment in different languages, and even use different scripts for the *same* language. We call such comments as **multi-glyphic comments**. For example, in Dainik Jagran, a popular Hindi newspaper, we observe that around 19% of the comments are in English, 34% are in Hindi, and more than 46% are in romanized Hindi (*Rohin*).

Figure 1 shows examples of comments made on a Hindi news article, highlighting the topical nuances and the problem of romanized text. As far as we know, there are no freely available machine translation systems that can handle romanized texts for the languages considered in this work. Due to lack of labeled data, supervised methods are hard to apply. In the recent past, it has been observed that topic models can be useful for modeling correspondence between article and comments (Das, Bansal, and Bhattacharyya 2014). In this paper we explore a hierarchical Bayesian approach for modeling multi-glyphic correspondence.

Contributions. In this paper, we develop the notion of *multi-glyphic correspondence*, i.e. comments in multiple languages and multiple scripts relating to an article in a single language and single script. To model *specific* correspondence, we apply multiple topic vectors (MTV) with a stick-breaking prior (SBP) following Das, Bansal, and Bhattacharyya (2014). The challenge in modeling *topical* correspondence across languages/scripts is that the source of multi-linguality is only through comments which are small, and noisy. Existing models assume that the proportion over topics is the same for all comments on an article, which is hardly true and conflicts with the modeling choice of MTV. We address this issue for some languages by incorporating an additional multi-lingual comparable corpus. When such corpora are not available (e.g. for romanized text), we show that introducing model sparsity helps. To address irrelevant comments, we use two types of correspondence: *global correspondence* when comments relate to a global topic outside the article, and *null correspondence* when the comment is not related to any news article at all. Thus the complete model addresses multi-glyphic comment correspondence and topical correspondence in a unified man-

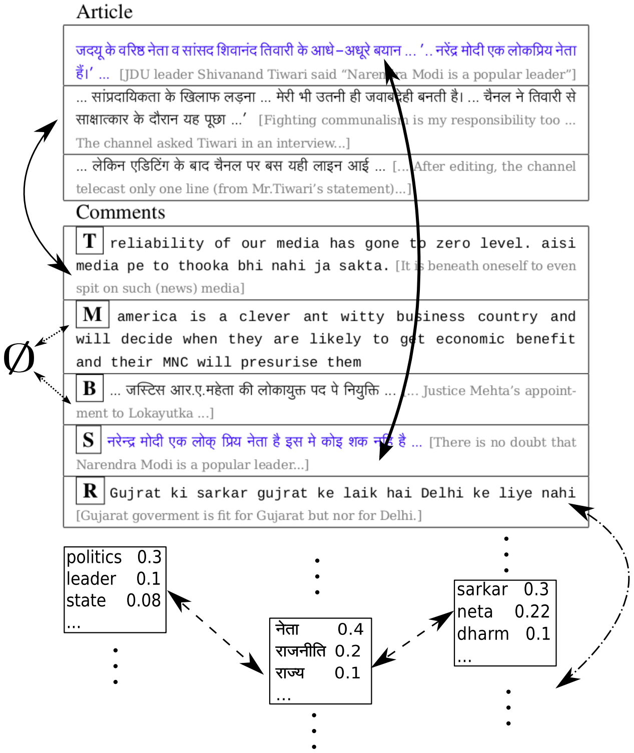

Article
जदयू के वरिष्ठ नेता व सांसद शिवानंद तिवारी के आधे–अधूरे बयान ... '.. नरेंद्र मोदी एक लोकप्रिय नेता हैं।' ... [JDU leader Shivanand Tiwari said "Narendra Modi is a popular leader"]
... सांप्रदायिकता के खिलाफ लड़ना ... मेरी भी उतनी ही जवाबदेही बनती है। ... चैनल ने तिवारी से साक्षात्कार के दौरान यह पूछा ...' [Fighting communalism is my responsibility too ... The channel asked Tiwari in an interview...]
... लेकिन एडिटिंग के बाद चैनल पर बस यही लाइन आई ... [... After editing, the channel telecast only one line (from Mr.Tiwari's statement)...]
Comments
T reliability of our media has gone to zero level. aisi media pe to thooka bhi nahi ja sakta. [It is beneath oneself to even spit on such (news) media]
M america is a clever ant witty business country and will decide when they are likely to get economic benefit and their MNC will presurise them
B ... जस्टिस आर.ए.महेता की लोकायुक्त पद पे नियुक्ति ... [... Justice Mehta's appointment to Lokayutka ...]
S नरेन्द्र मोदी एक लोक् प्रिय नेता है इस मे कोइ शक नहि है ... [There is no doubt that Narendra Modi is a popular leader...]
R Gujrat ki sarkar gujrat ke laik hai Delhi ke liye nahi [Gujarat goverment is fit for Gujarat but nor for Delhi.]
∅
politics 0.3
leader 0.1
state 0.08
...
नेता 0.4
राजनीति 0.2
राज्य 0.1
...
sarkar 0.3
neta 0.22
dharm 0.1
...

Figure 1: A Hindi article from Dainik Jagran with comments in English, Hindi, and *Rohin*. 5 types of comments are shown: topical (T), specific (S), corpus-topical (R), comment-topical (M), robotic (B). The first article segment, and the specific comment that discusses it, are marked in blue. English gloss for Hindi and *Rohin* text is shown next to it in gray. The different kinds of correspondence are shown using arrows: general (normal), specific (thick), global (dash-dotted), null (dotted), topical (dashed).

ner, where the two cooperate and reinforce the learning of each other. We name this model the **Multi-glyphic Correspondence Topic Model** (MCTM). We propose a collapsed Gibbs sampling inference procedure and evaluate the model on real-world data. We created manually annotated data sets for the comment classification task. The data and code have been released for public use (Tholpadi et al. 2014).

2 The Problem of Multi-glyphic Correspondence

We first discuss an empirical study on a real-life dataset. Then, we formally describe the problem objective.

Empirical Study

We performed an empirical study of comments for articles from two online sources—Dainik Jagran and Kendasampige[1]. Dainik Jagran (DJ) is India's highest–readership Hindi newspaper (MRUC 2013), while Kendasampige (KS) is a very popular online Kannada magazine (TOI 2011).

Dainik Jagran	#comments	#comments per article	#words per comment
English	3017 (19.4%)	2.6 ± 3.1	36.3 ± 34.8
Hindi	5358 (34.5%)	4.7 ± 4.3	45.7 ± 29.1
Rohin	7164 (46.1%)	6.3 ± 7.3	38.7 ± 26.2
Kendasampige	#comments	#comments per article	#words per comment
English	9188 (24.2%)	3.2 ± 4.1	27.0 ± 35.0
Kannada	20169 (53.0%)	6.9 ± 7.3	32.2 ± 24.3
Rokan	8679 (22.8%)	3.0 ± 2.7	12.0 ± 15.3

Table 1: Comment statistics for the data sets. (Columns 3 and 4 are averages.)

On both sites, readers were found to comment in the vernacular language (Hindi or Kannada), in English, and in *romanized* vernacular[2]. The romanized vernacular text entered by these users follow none of the standard romanization rules/systems[3]. As far as we know, there is no way to convert them into meaningful vernacular text. We will refer to romanized Hindi and Kannada as ***Rohin*** and ***Rokan***. Note that there exist no machine translation systems for *Rohin* and *Rokan*, so that existing methods for article-comment analysis *cannot* be used in conjunction with machine translation. Some statistics for the data sets are shown in Table 1. We find that romanized text constitutes a significant portion of the comments (46.1% and 22.8%). This motivates the need for methods that can handle romanized comments.

We analyzed a total of 300 articles from both data sets (Section 5.4). We observed users commenting in different languages and using different scripts. We also found that not all comments were related to the article, but often referred to topics from other articles or even extraneous themes. Based on our analysis, we defined the following **topical categorization** of comments (the numbers in brackets indicate the percentage of comments of that type in the analyzed data):

- **Topical** (37%): discusses the topic of the article. Some of the topical comments (43%) are **specific** comments—relevant to a specific segment of the article.
- **Corpus-topical** (17%): discusses topics that occur in other articles in the corpus. Typically, this happens when a commenter raises other issues that she thinks are relevant to this article.
- **Comment-topical** (40%): discusses topics that do not occur in any article in the corpus, e.g. compliments, expletives, personal/extraneous information, URLs, etc.
- **Robotic** (6%): appear almost verbatim in many articles, irrespective of topicality to the article (suggesting that it may have been posted by a program such as a web robot).

Figure 1 shows an example for each of the above categories.

[1] www.jagran.com, and www.kendasampige.com

[2] A vernacular comment written using the Latin script.

[3] See http://en.wikipedia.org/wiki/Romanization.

Definitions.
Romanized text: Text in Roman script in a language usually written using another script.
Multi-glyphic comments: A set of comments where two comments using the same script may be in different languages, and two comments using different scripts may be in the same language.
Multi-glyphic correspondence: The topical relationship between a news article and its set of multi-glyphic comments.
Dialect: a script–language pair, e.g. Devanagari Hindi, romanized Hindi, and (romanized) English are three dialects.

Input. We formally represent the dataset as follows. We are given a set of articles $\{w_d\}_{d=1}^{D}$. Each article consists of segments (e.g. paragraphs) $\{w_{ds}\}_{s=1}^{S_d}$. Each segment consists of words $\{w_{dsn}\}_{n=1}^{N_{ds}}$. Each article has a set of multi-glyphic comments x_d. The comments are in L dialects, and we group the comments in dialect l in the set $x_{dl} = \{x_{dlc}\}_{c=1}^{C_{dl}}$ for $l = 1 \ldots L$. A single comment x_{dlc} consists of the words $\{x_{dlcm}\}_{m=1}^{M_{dlc}}$.

Objective. Our objective is to develop a hierarchical Bayesian model suitable for news articles and multi-glyphic comments. In the literature, this kind of model is called a *correspondence* model. Here, $\{w_d\}$ are independent variables and $\{x_d\}$ are dependent variables. The novelty in this problem is that the two variables can be in different spaces (dialects).

Related Work

As far as we know, all previous work on news and comments has focused on comments in the article dialect. Kant, Sengamedu, and Kumar (2012) detect spam in comments. Mahajan et al. (2012) use features based on Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) to predict comment ratings. Ma et al. (2012) and Das, Bansal, and Bhattacharyya (2014) construct topic models for jointly modeling news and comments, similar to our work but in the monolingual setting. Sil, Sengamedu, and Bhattacharyya (2011) proposed a supervised approach to associate comments with segments. Our approach is unsupervised, and easier to adopt.

3 Modeling Multi-glyphic Correspondence

We first describe the basic model for correspondence and then propose our multi-glyphic correspondence topic model (MCTM).

Notation. We use the following notation in subsequent sections. Dir(), Categ(), Beta(), and Bern() represent the Dirichlet, Categorical, Beta, Bernoulli distributions, while SBP() represents a stick-breaking process (Ishwaran and James 2001). Unif(z) is a distribution assigning uniform probability to each component of z. We use $v \sim P$ to say that we sample a value for variable v from distribution P. $[N]$ is the set $\{1, \ldots, N\}$. V_l is the vocabulary size of dialect l. K and J are the number of article and comment topics, respectively. z and y denote the topic assignments for article and comment words, respectively. θ is the set of topic vectors for an article. ϕ denotes the word distributions for article topics.

Correspondence Topic Model. Correspondence LDA (CorrLDA) (Blei and Jordan 2003) can model correspondence between news articles and comments. Formally, the generative story is as follows.[4] First, sample the topics $\phi_k \sim$ Dir(β), $k \in [K]$. Then, for each article w_d:

- Sample $\theta \sim$ Dir(α).
- For each article word $w_n, n \in [N]$:
  - Sample $z_n \sim$ Categ(θ).
  - Sample $w_n \sim$ Categ(ϕ_{z_n}).
- For each comment word $x_{cm}, c \in [C],\ m \in [M_c]$:
  - Sample $y_{cm} \sim$ Unif(z).
  - Sample $x_{cm} \sim$ Categ($\phi_{y_{cm}}$).

3.1 Multi-glyphic Correspondence Topic Model

CorrLDA works only for the monolingual case, and is not directly applicable to even multi-lingual comments. If machine translation is available for the dialect of the comment, then CorrLDA can be applied. We can also apply SCTM for modeling specific comments (Das, Bansal, and Bhattacharyya 2014). However, machine translation is not available for many languages and romanized texts. This makes the problem of multi-glyphic comments beyond the scope of the state-of-the-art.

Multi-glyphic Correspondence (MCTM-D). We use a set of topics $\{\{\phi_{lj}\}_{j=1}^{K}\}$ for each dialect l to generate words in dialect l. Additionally, we assume that each topic in dialect l has a corresponding topic in all other dialects. Let the comments x for an article w be divided into subsets based on dialect, so that $x = \{\{\{x_{lcm}\}_{m=1}^{M_{lc}}\}_{c=1}^{C_l}\}_{l=1}^{L}$. The generative story for comments is modified to sample each $x_{lcm} \sim$ Categ($\phi_{ly_{lcm}}$) where $\phi_{lk} \sim$ Dir(β) is a topic in dialect l. Note that the topics ϕ_{lk} and $\phi_{l'k}$ *correspond*, i.e. they represent what people are saying in different dialects on the same topic.

Applying Multiple Topic Vectors (MCTM-DS). Das, Bansal, and Bhattacharyya (2014) pointed out the unsuitability of CorrLDA for capturing specific correspondence and proposed the specific correspondence topic model (SCTM) for this purpose. Following SCTM, we use *multiple topic vectors* (MTV) per article, and a *stick-breaking prior* (SBP) for the distribution over the topic vectors to model specific correspondence. Let $\{\theta_t\}_{t=1}^{T}$ be the set of topic vectors. To generate each word, we first sample one of the topic vectors, then the topic, and finally the word. Due to MTV,

[4] For ease of exposition, we will reuse variables at different granularities. The meaning should be clear from the context.

topic proportions vary across segments in an article. Thus we are able to model a *Rohin* comment relating to a particular paragraph of a Hindi news article.

Incorporating Topic Correspondence across Dialects. One key assumption made earlier is that each topic in a dialect has a corresponding topic in every other dialect. This in turn assumes that the distribution over topics is the same for all the comments as well as the article. However, in MCTM-DS, we vary the topic distribution across segments in an article. Moreover, in practice, we find that comments are small, noisy and vary heavily in vocabulary. For example, if a Hindi article has no or few topical English comments, the English topic corresponding to the article's Hindi topic would be of low quality. Thus the notions of multi-glyphicity and MTV apparently conflict and pose a significant hurdle.

We address this issue (MCTM-DSC) by using additional multi-lingual *comparable* corpora, and modeling topic correspondence similar to the polylingual topic model (Mimno et al. 2009). The use of such corpora achieves two purposes simultaneously: (1) improving topic quality in each dialect, and (2) improving topic correspondence across dialects.

Modeling Non-article Correspondence. The model developed so far performs quite well for modeling multi-glyphic correspondence. However, we observe that there are many *irrelevant* comments in the data—comments that do not correspond to the article. We make the following modeling choices for non-article correspondence.

Global correspondence (MCTM-DSG). To model the topicality of comments to other news articles in the corpus, we relax the constraint that the article is the source of all the topics in a comment, and allow the *article corpus* (the set of all articles in the corpus) to be a secondary source of topics. We introduce a *topic source distribution* p_{lc} for each comment, and a *topic source variable* q_{lcm} for each comment word. Each comment x_{lc} is now generated as follows.

- Sample $p_{lc} \sim$ Dir(λ).
- Sample $\rho_{lc} \sim$ Dir(δ).
- Sample $\epsilon_{lc} \sim$ Dir(η).
- For each word $x_{lcm},\ m \in [M_{lc}]$:
  - Sample $q_{lcm} \sim$ Categ(p_{lc}).
  - If $q_{lcm} = 1$:
    - Sample $b_{lcm} \sim$ Categ(ρ_{lc}).
    - Sample $y_{lcm} \sim$ Unif$(z_{b_{lcm}})$.
  - Else If $q_{lcm} = 2$: Sample $y_{lcm} \sim$ Categ(ϵ_{lc}).
  - Sample $x_{lcm} \sim$ Categ$(\phi_{ly_{lcm}})$.

Null correspondence (MCTM-DSGN). To model topics that occur in comments but not in the articles, we extend MCTM to incorporate the *comment corpus* (the set of all comments) as another topic source. We introduce a secondary set of topics $\psi_{lj},\ l \in [L], j \in [J]$ that we call *comment topics*. These word distributions are used only for generating comments. We add a component to p_{lc} where p_{lc3}=Pr[comment corpus is a topic source for a comment word], and use a *comment corpus topic vector* χ_{lc}, a distribution over comment topics. The generative story is similar to MCTM-DSG, with the following differences: (1) For each comment, we also sample $\chi_{lc} \sim$ Dir(ξ). (2) If $q_{lcm} = 3$, we sample $y_{lcm} \sim$ Categ(χ_{lc}), and $x_{lcm} \sim$ Categ$(\psi_{ly_{lcm}})$.

Roboticity. Interestingly, the model associates robotic comments with comment topics, even when there are articles in the corpus that may seem topically related to these comments. This is because each robotic comment occurs many times in the corpus, and thus the probability of generating them from comment topics is generally greater. Thus the MCTM-DSGN model captures both comment-topical and robotic comments (but fails to distinguish between the two).

Incorporating Sparsity. We describe two kinds of sparsity and propose a joint sparsity model.

Topic sparsity. Wang and Blei (2009) force each topic to be a distribution over a small subset of the vocabulary to get sparse topics without sacrificing smoothness. For this, we define *topic sparsity parameters* κ_{lk}=fraction of the vocabulary included in topic ϕ_{lk}, and binary *sparsity variables* $a_{lkv} = 1$ if $v \in \phi_{lk}$, and 0 otherwise. The generative story for each topic $\phi_{lk},\ l \in [L], k \in [K]$ becomes:

- Sample $\kappa_{lk} \sim$ Beta(ν).
- For each word $v \in [V_l]$: Sample $a_{lkv} \sim$ Bern(κ_{lk}).
- Sample $\phi_{lk} \sim$ Dir(βa_{lk}).

Here, Dir(βa_{lk}) is a distribution over the subset of the vocabulary defined by a_{lk}.

Word sparsity. Das, Bansal, and Bhattacharyya (2014) introduce *word sparsity* (they call it "topic diversity") by forcing each word to belong to a small subset of the topics. For this, we define *word sparsity parameters* σ_{lv}=fraction of the K topics that word v may belong to. Using a_{lkv} as before, the generative story changes to:

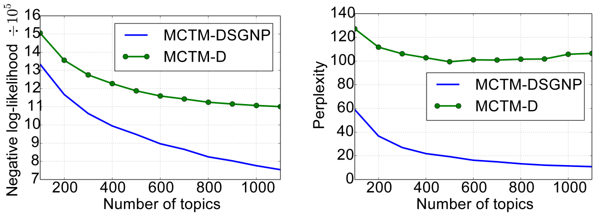
- For each dialect $l \in [L]$, for each word $v \in [V_l]$: Sample $\sigma_{lv} \sim$ Beta(μ).
- For each topic $\phi_{lk},\ l \in [L], k \in [K]$:
  - For each word $v \in [V_l]$: Sample $a_{lkv} \sim$ Bern(σ_{lv}).
  - Sample $\phi_{lk} \sim$ Dir(βa_{lk}).

Joint topic-word sparsity (MCTM-DSGNP). We combine the benefits of both schemes by using a **joint topic-word sparsity** scheme. The steps of both generative stories are combined, but with one modification—we now sample $a_{lkv} \sim$ Bern$(\kappa_{lk}\sigma_{lv})$. However, this causes coupling between κ_{lk} and σ_{lv}, making it difficult to integrate them out for doing collapsed Gibbs sampling. To decouple κ_{lk} and σ_{lv}, we introduce auxiliary *topic sparsity variables* e_{lkv} and *word sparsity variables* f_{lkv}, and define $a_{lkv} = e_{lkv} f_{lkv}$.

For the complete plate diagram and generative process, see the supplementary material (Tholpadi et al. 2014).

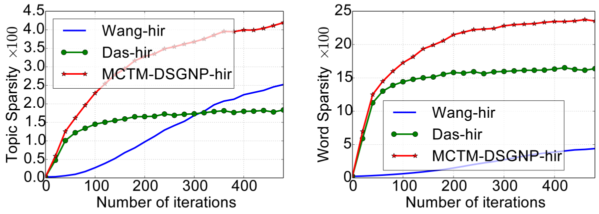
4 Collapsed-Blocked Gibbs Sampling

We use Gibbs sampling for estimating the MCTM model. We want to **collapse** all real-valued variables and preserve only categorical variables so that we can converge faster,



(a) Training data fit (b) Held-out data perplexity

Figure 2: Model fit for the DJ corpus (lower is better).



(a) Topic sparsity (b) Word sparsity

Figure 3: Sparsity levels achieved by different models on the DJ corpus for *Rohin* (higher is better).

and also to avoid round-off errors. The main challenges that make the inference non-standard are:

- Coupling between κ and σ—we solve this by introducing auxiliary variables e and f.
- Sampling r (due to the SBP prior)—we use the equivalence of the SBP to the Generalized Dirichlet (GD) distribution (Connor and Mosimann 1969) and derive the sampling update.
- Interdependence between q, b, y, and between a, e, f—we handle this issue by doing **blocked** Gibbs sampling.

The details of the inference procedure are given in the supplementary material (Tholpadi et al. 2014).

5 Experiments

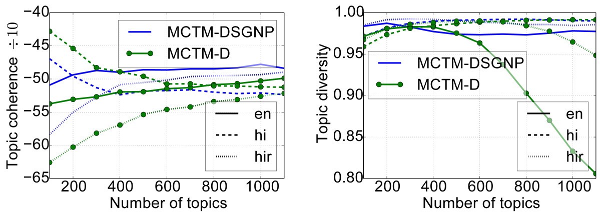
We evaluated the MCTM model on two data sets crawled from the web. Since previous work are not applicable to our setting, we used the MCTM-D model as the closest extension to existing methods and compare it with all model variants. We approached the evaluation from three angles:

- Is the model a good fit for the data?
- How does sparsity affect topic quality?
- Can the model detect comment categories?

5.1 Data and Preprocessing

We gathered articles with comments from a Hindi newspaper Dainik Jagran (DJ) (5699 articles) and from a Kannada magazine Kendasampige (KS) (5329 articles). For each article, we extracted segments that approximately corresponded to paragraphs. We identified the language of the comments using the Unicode code block or, in the case of *Rohin*/*Rokan*, a language detection library (Shuyo 2010). We selected articles with at least 5 comments and constructed 2 data sets with 1142 (DJ) and 2905 (KS) articles. Section 2 discusses some statistics of the data sets. The vernacular text was stemmed (Reddy and Sharoff 2011) and normalized to map variant Unicode sequences of a word to a unique sequence. We also hand-crafted stop word lists for each of the 5 dialects (totalling 4368 words).

Language	English	Hindi	Kannada	*Rohin*	*Rokan*
# stop words	627	1118	595	1000	1028
V_l	∼6K	∼10K	∼15K	∼6K	∼2.5K



(a) Topic coherence (b) Topic diversity

Figure 4: Topic quality (for the top 20 words) achieved on the DJ corpus (higher is better).

For the specific comment detection task, we used a multilingual comparable corpus of 1000 document pairs extracted from the English and Kannada Wikipedias.

5.2 MCTM-DSGNP is a Better Fit for Data

We want to evaluate whether the model is able to learn well, and whether it is a good fit for the data. We ran the model on the DJ corpus with the following configuration: T=$\frac{S_{dl}}{2}$, K=500, J=4, α=.01, β=.01, ω=.01, γ=[.05,.1], $\delta \propto (N_{dls})_{s=1}^{S_{dl}}$, η=.01, ξ=.01, $\lambda \propto$[.72,.18,.1]×$M_{dll'c}$. We did a grid search for K and J and chose values that gave the best human-readable topics. T was set such that there was one topic vector for every two segments. δ was set so that larger segments were more likely to generate comments. The λ used captures our rough guess that around 10% of the comments are comment-topical or robotic, less than 20% are corpus-topical, and the remaining are relevant to the article. The other parameters were chosen to encourage peaked distributions for topics and words (Heinrich 2009).

We found that both MCTM-D and MCTM-DSGNP have good convergence, and the training data likelihood stabilizes at around 500 iterations. We evaluated how well the model fits the training data and held-out test data. Figure 2(a) shows the training data negative log-likelihood for different values of K for both models. MCTM-DSGNP clearly outperforms the baseline, and does even better at higher K. To make sure the model does not overfit, we computed the perplexity on held-out data (Figure 2(b)), and found that MCTM-DSGNP can handle unseen data better than MCTM-D.

5.3 Good Topic Quality with High Sparsity

Wang and Blei (2009) define the *complexity* of a topic ϕ_{lk} as the number of words that belong to the topic, i.e.

Kendasampige	Specific (S)		
Model	P	R	F1
MCTM-D	.188	.875	.401
MCTM-DS	.186	**.890**	.400
MCTM-DSG	.189	.874	.402
MCTM-DSGN	.193	.835	.405
MCTM-DSGNP	.197	.768	.395
MCTM-DSGNPC	**.217**	.706	**.405**

Dainik Jagran	**Corpus+Comment +Robotic (RMB)**			**Comment+ Robotic (MB)**			**Robotic (B)**		
Model	P	R	F1	P	R	F1	P	R	F1
MCTM-D	.738	.303	.414	.584	.343	.418	.051	.116	.122
MCTM-DS	.765	.314	.430	.617	**.351**	.438	.123	.420	.353
MCTM-DSG	.759	.306	.420	.614	.335	.425	.137	.393	.367
MCTM-DSGN	**.825**	.413	.538	.788	.262	.480	.325	**.879**	.828
MCTM-DSGNP	.762	**.515**	**.586**	**.776**	.204	**.485**	**.438**	.862	**.841**

Correspondence types: **D** multi-glyphic, **S** specific, **G** global, **N** null, **P** sparsity, **C** comparable corpora

Table 2: Precision (P), recall (R) and F1-score (F1) for identifying different kinds of comments.

complexity$_{lk} = \sum_v a_{lkv}$. Since we defined the model objectives in terms of sparsity, we defined two *sparsity* metrics to measure the performance of the sparsity schemes:

$$\text{topic sparsity}_{lk} = \frac{1}{\sum_v a_{lkv}}\ ,\ \text{word sparsity}_{lv} = \frac{1}{\sum_k a_{lkv}}.$$

Figure 3 shows the average sparsity (over all topics/words) for *Rohin* on the DJ corpus (the results were similar for other dialects). We compare the topic, word, and joint sparsity schemes. We see that our scheme achieves sparsity very early, and leads to higher sparsity.

A natural question to ask is: Does the high topic sparsity affect topic quality? To check this, we plot topic coherence (Mimno et al. 2011) and topic diversity (Das, Bansal, and Bhattacharyya 2014) for the different dialects and compare it with MCTM-D which used no sparsity schemes (Figure 4). We see that there is no loss in topic quality or diversity in spite of the high sparsity. The comment topics discovered were especially interesting, since we found that they could be categorized into different classes such as “compliments”, “foreign words”, “commenter names”, “expletives”, etc. (Tholpadi et al. 2014).

5.4 Comment Category Detection Task

We apply the different model variants to the task of detecting different kinds of comments in articles. For the purpose of evaluation, we created two gold standard data sets by manual annotation.

Specific correspondence Data Set. We annotated 202 articles in the KS corpus, together containing 6192 comments, of which 3075 were in *concomitant* dialects, i.e. comment dialects other than the article language. For each article, we asked an annotator to read the article body, and then annotate each *concomitant dialect* comment[5] as ‘Non-Topical’, ‘Topical, but not specific’, or ‘Topical, and specific to segments $s_1, s_2, \ldots$’, where s_i is a segment index.

Topical Correspondence Data Set. We annotated 102 articles in the DJ corpus, together containing 1379 comments, of which 855 were in concomitant dialects. Each comment was marked as ‘Topical’, ‘Corpus-topical’, ‘Comment-topical’, or ‘Robotic’.

Algorithms for Comment Detection. As far as we know, there are no freely available translation/transliteration systems or parallel/comparable corpora in *Rohin*/*Rokan*. Hence, none of the existing cross-language methods for classification can be used as baselines for our data sets. Given this constraint, we came up with the best possible baseline for our setting, viz. the MCTM-D model.

For the MCTM-DSGNP model, we used a combination of the topic sources for a comment ($b_{dll'c}$) and the topic vector to determine the category of the comment. For the MCTM-D model, we constructed topic vectors for each comment and segment, and used cosine similarity with tuned thresholds to determine the comment category. The details of the algorithms are given in the supplementary material (Tholpadi et al. 2014).

Results. The results of the evaluation on Kendasampige and Dainik Jagran are shown in Table 2. For the specific comment detection task, we see a steady improvement in precision (up to 15%), especially by using comparable corpora (MCTM-DSGNPC). Also note that the MCTM-D method requires tuning thresholds to achieve the best performance, while the MCTM-DSGNPC method requires no tuning, which is useful when labeled data is not available.

For detecting irrelevant comments (**RMB** and **MB**), MCTM-DSGN gives huge gains over the simpler models (up to **42%**). This is expected since this model explicitly captures global and null correspondence. In particular, we see massive improvement (up to **589%**) on the robotic comment detection task. Also observe that introducing sparsity (MCTM-DSGNP) almost always helps improve performance on all tasks.

6 Conclusion

In this paper, we studied the phenomenon of multi-glyphic comments to online news, especially the presence of *romanized* text, and identified challenges in learning different kinds of topical correspondence. We developed the MCTM model to address the challenges using a hierarchical Bayesian approach. Evaluation on real-world data sets show the efficacy of the model, and potential for various applications. To facilitate further research in this new area, we have released the annotated data sets and code for public use.

Acknowledgments

We thank Adway Mitra for useful discussions, and Chaitra Shankar for help with the annotation. This work is supported by the Dept. of Science and Technology, Govt. of India.

[5] Analyzing comments in the article dialect has been addressed in previous work; we focus on concomitant dialects in this work.

References

Blei, D. M., and Jordan, M. I. 2003. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 127–134. ACM.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3:993–1022.

Connor, R. J., and Mosimann, J. E. 1969. Concepts of independence for proportions with a generalization of the dirichlet distribution. *Journal of the American Statistical Association* 64(325):194–206.

Das, M. K.; Bansal, T.; and Bhattacharyya, C. 2014. Going beyond corr-lda for detecting specific comments on news & blogs. In *Proceedings of the 7th ACM international conference on Web search and data mining*, 483–492. ACM.

Heinrich, G. 2009. Parameter estimation for text analysis. Technical report.

Ishwaran, H., and James, L. F. 2001. Gibbs sampling methods for stick-breaking priors. *JASA* 96(453).

Kant, R.; Sengamedu, S. H.; and Kumar, K. S. 2012. Comment spam detection by sequence mining. In *Proceedings of the fifth ACM international conference on Web search and data mining*, 183–192. ACM.

Ma, Z.; Sun, A.; Yuan, Q.; and Cong, G. 2012. Topic-driven reader comments summarization. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, 265–274. ACM.

Mahajan, D. K.; Rastogi, R.; Tiwari, C.; and Mitra, A. 2012. Logucb: an explore-exploit algorithm for comments recommendation. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, 6–15. ACM.

Mimno, D.; Wallach, H. M.; Naradowsky, J.; Smith, D. A.; and McCallum, A. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, 880–889. Association for Computational Linguistics.

Mimno, D.; Wallach, H. M.; Talley, E.; Leenders, M.; and McCallum, A. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262–272. Association for Computational Linguistics.

MRUC. 2013. Indian Readership Survey. http://mruc.net/sites/default/files/irs_2013_topline_findings.pdf. [Online; accessed 12-Nov-2014].

Reddy, S., and Sharoff, S. 2011. Cross language pos taggers (and other tools) for indian languages: An experiment with kannada using telugu resources. In *Proceedings of the Fifth International Workshop On Cross Lingual Information Access*, 11–19. Chiang Mai, Thailand: Asian Federation of Natural Language Processing.

Shuyo, N. 2010. Language detection library for java.

Sil, D. K.; Sengamedu, S. H.; and Bhattacharyya, C. 2011. Supervised matching of comments with news article segments. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2125–2128. ACM.

Tholpadi, G.; Das, M.; Bansal, T.; and Bhattacharyya, C. 2014. Supplementary material to Relating Romanized Comments to News Articles by Inferring Multi-glyphic Topical Correspondence. http://mllab.csa.iisc.ernet.in/mctm. [Online; accessed 12-Nov-2014].

TOI. 2011. Kannada Sahitya Sammelana: Kannada blogs, sites give the write connect. http://timesofindia.indiatimes.com/articleshow/7434091.cms. [Online; accessed 12-Nov-2014].

Wang, C., and Blei, D. M. 2009. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In *Advances in neural information processing systems*, 1982–1989.

Wikipedia. 2014. Languages of India (Official languages). http://en.wikipedia.org/wiki/Languages_of_India#Official_languages. [Online; accessed 18-Nov-2014].