Cluster Labeling for Multilingual Scatter/Gather using Comparable Corpora

Goutham Tholpadi, Mrinal Kanti Das,
Chiranjib Bhattacharyya, and Shirish Shevade

Computer Science and Automation, Indian Institute of Science, Bangalore, India
{gtholpadi,mrinal,chiru,shirish}@csa.iisc.ernet.in
http://csa.iisc.ernet.in

Abstract. Scatter/Gather systems are increasingly becoming useful in browsing document corpora. Usability of the present-day systems are restricted to monolingual corpora, and their methods for clustering and labeling do not easily extend to the multilingual setting, especially in the absence of dictionaries/machine translation. In this paper, we study the *cluster labeling* problem for *multilingual* corpora in the absence of machine translation, but using comparable corpora. Using a variational approach, we show that multilingual topic models can effectively handle the cluster labeling problem, which in turn allows us to design a novel Scatter/Gather system *ShoBha*. Experimental results on three datasets, namely the Canadian Hansards corpus, the entire overlapping Wikipedia of English, Hindi and Bengali articles, and a trilingual news corpus containing 41,000 articles, confirm the utility of the proposed system.

Keywords: cluster labeling, multilingual, Scatter/Gather, comparable corpora, topic models.

1 Introduction

Over the last decade Scatter/Gather-based systems [7] have emerged as useful tools for browsing document corpora. The core of the system relies on the ability to automatically form and label clusters. Cluster labels are short key phrases/words that serve as guides for browsing. Previous work [15, 3] has shown that cluster-based browsing improves information access. Industrial adoption[1] and recognition[2] have demonstrated its usefulness. Cluster-based browsing systems such as Scatter/Gather have emerged as a tool of choice for monolingual text corpora but their applicability to multilingual corpora is extremely limited. In this paper we study the problem of designing Scatter/Gather systems for multilingual corpora[3].

[1] Vivisimo, Carrot[2], Eigencluster, Clusty, Yippy, etc.

[2] Vivisimo won the "best meta-search engine award" by SearchEngineWatch.com from 2001 to 2003.

[3] A multilingual corpus is one where each document is in a single language, but there are documents in many languages.

The success of the method is crucially dependent on solving the problem of automatic clustering and the subsequent labeling of clusters. Existing methods [4, 12, 10] use accurate dictionaries or powerful machine translation for labeling multilingual corpora. When machine translation is unavailable, both problems—*clustering a multilingual corpus* and *labeling a multilingual cluster*—become challenging.

Contributions. The main contribution of the paper is a novel labeling procedure which is suited for multilingual document clusters. The procedure, which uses comparable corpora[4], uses an adaptation of multilingual topic models where the computationally intensive Gibbs sampling step (as described in [14]) is replaced by a variational inference procedure. To evaluate the efficacy of the method, we built a system called *ShoBha* to do multilingual Scatter/Gather[5]. We evaluate the labeling quality on several real-world data sets including the Canadian Hansards and a Wikipedia data set culled from the English, Hindi and Bengali Wikipedias. We observe that *ShoBha* performs reasonably well when compared to the gold standard. We report the results of a user study and show that the model works reasonably well in a practical Scatter/Gather setting. We built language resources for our experiments, and have released them for public use.[6]

The paper is organized is as follows. First we describe the problem, the topic model and the multilingual Scatter/Gather system in Sect. 2. We discuss the evaluation method and experimental results in Sect. 3. We mention the related work in Sect. 4 and then conclude.

2 Multilingual Labeling using Topic Models

Problem Description. Given a document collection $\mathcal{C} = \{d_i\}_{i=1}^{D}$ where each document d_i is known to be in language $l_i \in \{1, \ldots, L\}$, and L is the number of languages, the objective is to cluster $\mathcal{C}$ into clusters $\{\mathcal{C}_j\}_{j=1}^{C}$ and generate label sets $\{\{S_j^l\}_{l=1}^{L}\}_{j=1}^{C}$ that summarize each cluster in every language. Here $S_j^l \in \{1, \ldots, V_l\}^k$ where V_l is the vocabulary size of language l, and k is the number of labels in a label set.

State-of-the-art monolingual ($L = 1$) labeling methods are based on word frequencies and do not work for multilingual clusters ($L \geq 2$). For example, if a cluster of 100 documents contains only 2 documents in one language, the labels generated in that language using monolingual methods can hardly represent the cluster as a whole.

[4] A comparable corpus in L languages can be organized into a set of tuples, where each tuple contains L documents—all on the same topic, but each in a different language. While building machine translation systems is hard, comparable corpora are relatively easy to obtain, especially from the Web (e.g. Wikipedia, news, multilingial websites, etc.)

[5] A Scatter/Gather system allows user to 'scatter' documents into a number of clusters, summarizing each cluster with keyword labels. The user then 'gather's interesting clusters and re-scatters them.

[6] `http://mllab.csa.iisc.ernet.in/shobha`

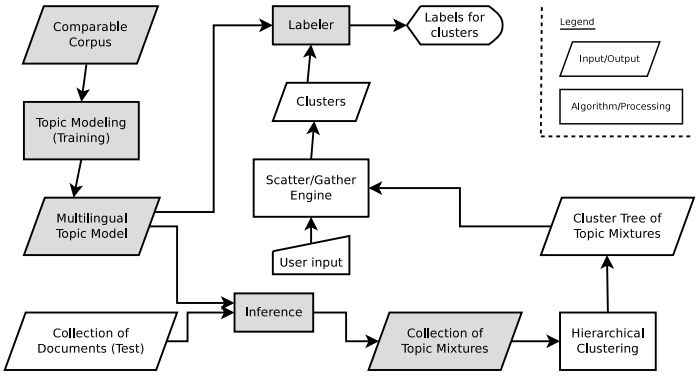

Fig. 1: Architecture of *ShoBha*, a system for multilingual Scatter/Gather. The shaded portions highlight what components are required over and above monolingual Scatter/Gather.

2.1 Scatter/Gather System Architecture

In order to enable multilingual Scatter/Gather, we need to solve both the clustering and the labeling problems mentioned in Sect. 1. We used topic models to solve both problems, but focus on the labeling problem in this paper. We built a Scatter/Gather system called *ShoBha* that incorporates our approach. Given a comparable corpus (and no other language resources), the system learns multilingual topics (top-left in Fig. 1). It takes as input any multilingual document collection that the user wants to browse (bottom-left in figure). It infers topic mixtures for all documents and then builds a cluster hierarchy using the topic mixtures as the features. The Scatter/Gather engine (similar to [11]) enables cluster-based browsing of the hierarchy, and provides labels on-the-fly for each cluster, in the user's preferred language.

2.2 Learning Multilingual Topics

We use the Polylingual Topic Model as described in [14] for learning multilingual topics. Instead of Gibbs sampling, we propose an efficient *variational approximation* method following [1]. This was more scalable than the sampling-based approach (as discussed below). First, we derive the updates for learning the model parameters, after defining notation.

Let V^l be the size of the vocabulary in language l. A document d in language l is a set of words $\{w_1, w_2, \ldots, w_N\}$ where w_n is a vector of size V^l with one component set to 1, and all the others set to 0. If $w_{ni} = 1$, then w_n represents the i^{th} word in the vocabulary. Given T topics, a topic mixture θ is a T-vector sampled from Dirichlet(α) and is used to generate d. z_n is the topic for the n^{th} word w_n in d. θ and z are the latent variables, and β_t is the word distribution for the t^{th} topic. The joint distribution of any document d after marginalizing over the latent variables is $p(d) = \int_\theta \sum_z p(\theta, z, d|\alpha, \beta) d\theta$.

The existing approach to solve this problem uses Gibbs sampling. Instead, we introduce free variational parameters $\phi^1, \phi^2, \ldots \phi^L$ and γ as shown in Fig. 2.

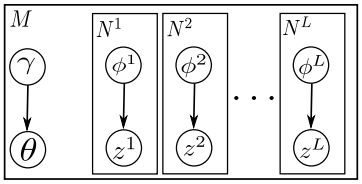

Fig. 2: Variational approach for the multilingual topic model. M is the number of document tuples.

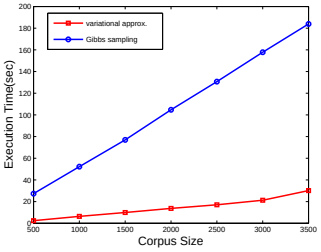

Fig. 3: Comparison of variational *vs.* sampling approaches.

Representing $\{\phi^1, \phi^2, \ldots \phi^L\}$ by ϕ, we define a variational distribution q as $q(\theta, z|\gamma, \phi) = q(\theta|\gamma) \prod_{l=1}^{L} \prod_{n=1}^{N^l} q(z_n^l|\phi_n^l)$ where γ is a Dirichlet parameter, and $\phi^1, \phi^2, \ldots, \phi^L$ are multinomial parameters, and N^l is the number of words in the document in language l. Following the approach in [1], we approximate the posterior joint distribution p of the latent variables by the factorizable variational distribution q and obtain an EM procedure. The update for α is as described in [1]. The other updates are[7]

$$\gamma_i = \alpha_i + \sum_{l=1}^{L} \sum_{n=1}^{N^l} \phi_{ni}^l \ , \qquad \beta_{ij}^l \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d^l} \phi_{dni}^l w_{dn}^{(l)j} \ ,$$

$$\text{and} \qquad \phi_{ni}^l \propto \beta_{i\nu}^l \exp\left(\Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{k} \gamma_j\right)\right) \ .$$

We tested the Gibbs sampling and the variational approximation approaches for scalability. We compared the execution time taken by the two methods to reach the same likelihood value. We repeated this for different corpus sizes (of the Bengali Wikipedia dataset). The variational approach was 87% faster, on average (Fig. 3), and was used for all further experiments.

Clustering using Topics. After learning the model parameters α and β, we infer topic mixtures for our target corpus. We infer a T-vector γ for each document, and define the T-vector η such that $\eta_i = \frac{\gamma_i}{\sum_j \gamma_j}$ as the topic mixture for that document. This maps documents in all the languages into a common space where Euclidean distance is a good approximation of semantic distance. We can now use any clustering algorithm on this document collection. We used *agglomerative average-linkage* clustering [13].

2.3 Labeling Multilingual Clusters

Given a multilingual cluster $\{d_1, \ldots, d_n\}$ and a multilingual topic model $\mathcal{M}$, we want to generate a label set $S = (w_1, \ldots, w_k)$ (a k-tuple of words) in language l. For this, we need to solve $\max_{S \in \{1,\ldots,V^l\}^k} p(S|d_1, \ldots, d_n, \mathcal{M})$. We first map all documents into a common semantic space by replacing each document (d_i) by

[7] Ψ is the digamma function.

its topic vector (η_i). The search space is exponential in k. If we assume that the words in S are independent, the problem becomes $\max_{w_i \in \{1,\ldots,V^l\}} p(w_i|\eta^1, \ldots, \eta^n, \mathcal{M})$ for $i = 1, \ldots, k$. Here, the posterior distribution p is unknown. If we assume that the documents are sampled independently and that their distribution is uniform, we can show that, for any word w,

$$p(w|\eta^1 \ldots \eta^n, \mathcal{M}) \propto p(w|\mathcal{M}) \prod_{j=1}^{n} \frac{p(w|\eta^j, \mathcal{M})}{p(w|\mathcal{M})}$$

where $p(w|\eta, \mathcal{M}) = \sum_{t=1}^{T} p(w|t, \mathcal{M}) p(t|\eta) = \sum_{t=1}^{T} \beta_{tw} \eta_t$, and β_{tw} is the w^{th} component of β_t. Since $\mathcal{M}$ depends on the training corpus $\mathcal{C}$, we assume that w is conditionally independent of $\mathcal{M}$ given $\mathcal{C}$. Thus we get $p(w|\mathcal{M}) = p(w|\mathcal{C}) = \frac{N_w}{\sum_{w'} N_{w'}}$, where N_w is the number of occurrences of w in $\mathcal{C}$.

Note that we compute a distribution (rather than an ordering) over words. The probabilities are useful for other methods that use "important" terms to enrich the set of candidate labels, and also for rendering in visualizations such as tag clouds (e.g. larger fonts for higher probabilities). This method can be used to generate labels in any of the L languages. For each cluster, generating the labels costs $O(V^l nT + V^l \log V^l)$. We observed that replacing the set of η's by the cluster centroid also gave good labels and reduced the cost to $O(V^l T + V^l \log V^l)$. This approximation was used in all further experiments and is evaluated in Sect. 3.

3 Experiments

In this section, we evaluate the quality of the labels generated, given a cluster[8]. Cluster labels are mainly for human consumption. Hence, the ideal method of evaluation is to ask human experts to carefully examine each cluster and judge label quality. Accordingly, we performed a user study on a news corpus (Sect. 3.6). But we first discuss quantitative evaluation. Since there were no freely available human-annotated labeling data for multilingual clusters, we developed an evaluation scheme that leveraged comparable corpora.

3.1 A Monolingual Baseline for Quantitative Evaluation

One popular labeling evaluation method uses human-annotated 'correct' labels for each cluster (the '*gold* standard'), and measures how close system-generated labels are to the gold standard. [2] used 20 News Groups data and the Open Directory Project as gold standard data—each newsgroup/hierarchy node is a cluster, and the newsgroup/node description is the gold standard label set. To the best of our knowledge, there do not exist freely available *multilingual* data sets of this nature.

State-of-the-art (SOTA) methods for labeling monolingual clusters are quite effective [2]. So, if any new labeling method generated labels that were close to the SOTA labels, this would indicate good quality. [9]

[8] The clustering itself is not evaluated; the clusters are assumed to be of good quality.

[9] Note that if the new labels were not close to the SOTA labels, it does not necessarily indicate poor quality.

Evaluation Setup. Suppose we have a multilingual cluster $\mathcal{C}$ (in L languages) having D documents. We translate each document into all the other languages. We get L monolingual clusters $\{\mathcal{C}_l\}_{l=1}^{L}$, each containing D documents.

We can arrive at this kind of setup in another way. Suppose we have a comparable corpus in L languages. Also suppose that we are able to cluster this corpus.[10] Choose any cluster. Let it contain D tuples. From this, it is easy to extract monolingual clusters $\{\mathcal{C}_l\}_{l=1}^{L}$ each containing D documents. Also, by randomly selecting a document from each tuple, we can construct a multilingual corpus $\mathcal{C}$.

We can use our method on $\mathcal{C}$ to generate labels in each of the L languages. We can use SOTA monolingual methods on $\{\mathcal{C}_l\}_{l=1}^{L}$ to generate labels in each language. If the labels generated by our method are close to the SOTA labels in a particular language, it indicates good labeling quality in that language. Instead of choosing $\mathcal{C}$ randomly, we used $\mathcal{C} = \mathcal{C}_l,\ l = 1 \ldots L$, i.e. we evaluated the method for each document language–label language pair.

A 'State-of-the-art' Monolingual Labeling Method. [2] describes several SOTA methods for monolingual cluster labeling. But most of them are slow—for each cluster, they require computations over the entire corpus. This is infeasible for our interactive system *ShoBha*, where labeling needs to be done on-the-fly whenever the user scatters or gathers clusters. The fastest of their SOTA methods was based on *ctf.cdf.idf* features, and this was used in our evaluation (as the '*gold* standard'). In this method, the terms with the highest values of *ctf.cdf.idf* in the centroid of the cluster are used as the labels. The value of *ctf.cdf.idf* of a term t with respect to a cluster C is calculated as

$$ctf.cdf.idf(t, C) = ctf(t, C).cdf(t, C).idf(t)$$

where $ctf(t, C) = \frac{1}{|C|} \sum_{d \in C} tf(t, d)$, $cdf(t, C) = log(n(t, C) + 1)$, $n(t, C)$ is the document frequency of t in C, and $tf(t, d)$ is the term frequency in d, and $idf(t)$ is the inverse document frequency of t in the entire corpus.

3.2 Evaluation Metric and Plots

We generated clusters for each test corpus and manually selected coherent clusters for measuring label quality. Given a language l, we generated the gold standard label sets using $\mathcal{C}_l$ (as defined in Sect. 3.1). Call these sets $\{G_i\}_{i=1}^{C}$ for C clusters. Then, given a method M, we generated the label sets $\{M_i\}_{i=1}^{C}$ using $\mathcal{C}_l$. The precision of the method M for language l is calculated as $\frac{1}{C} \sum_{i=1}^{C} \frac{|G_i \cap M_i|}{|M_i|}$. We plot the precision for different values of k (the number of labels) for each method, and show plots for each language.

For example, in Fig. 4 (left), we evaluate the English labels (l = EN) for an English-French corpus. For a cluster, we first generate the gold labels in English using the English-only cluster $\mathcal{C}_{\text{EN}}$. Next, we generate topic vectors from $\mathcal{C}_{\text{EN}}$ and generate English labels using our system. This is reported as EN-TM (labels generated from $\mathcal{C}_{\text{EN}}$ using the topic model). This measures the quality of the

[10] Imagine clustering on tuples of documents rather than single documents—since all documents in a tuple are related, they must belong to the same cluster.

English labels when all the documents in the cluster are in English. Next, we get topic vectors for the French-only cluster $\mathcal{C}_{\text{FR}}$ and generate *English* labels using our system. This is reported as FR-TM (labels generated from $\mathcal{C}_{\text{FR}}$ using the topic model). This measures the quality of the English labels when all the documents in the cluster are in French. The plots in Fig. 4 (right) and Fig. 5 should be interpreted in a similar manner.

3.3 Data Sets

Two comparable corpora (derived from the Canadian Hansards, and from Wikipedia) were used to quantitatively evaluate the method. A multilingual news corpus was used for qualitative evaluation.[11]

High-quality English-French Data (HANS). To accurately measure the efficacy of the method and compare its performance with the machine translation approach, we chose a very clean data set in English-French—the Canadian Hansards House Debates corpus. The training set has 948K sentence pairs spread over 313 document pairs. The sentences are too small to model documents; and each document is too big to be comparable. So, we split the documents to get 13,611 training and 964 test document pairs. After removing stopwords, the vocabulary was restricted to the top 10K words (ordered by *tf.idf*). The preprocessed corpus had 5M English and 7.1M French words.

Noisy English-Hindi-Bengali Data (WIKI). We chose the English, Hindi and Bengali to evaluate the method on a noisy corpus from a different language family[12]. We created an English-Hindi-Bengali comparable corpus from the intersection of the Wikipedias in the three languages (3349 document triplets).

After removing stopwords, we restricted the data to the top 5000 words in the vocabulary, ordered by term counts. Usually the most frequent words are chosen as labels; so the labeling quality is not affected significantly. After preprocessing, we obtained 2700 training triplets and 649 test triplets.

English-Hindi-Bengali News Corpus (FIRE). We use the FIRE news corpus[13] to demonstrate a real-world application of Scatter/Gather to browsing multilingual news. We used all news articles for the year 2004 from an English newspaper (14346 articles) and a Bengali newspaper (12359 articles), and 15000 articles from a Hindi newspaper [14]. Since the English and Bengali newspapers are based in the same city, and have similar timelines, it is likely that there are articles from both languages in any cluster. The data set was preprocessed using the same steps as the Wikipedia data.

[11] Note that the choice of languages was influenced by the availability of testing data, and not just resource scarcity. The methods applied did not use extra resources, in order to simulate resource scarcity.

[12] Hindi and Bengali are more agglutinative when compared to English and French.

[13] `http://www.isical.ac.in/~clia/`

[14] Information about the year of publishing was not available, but was known to be within the range 2004-07.

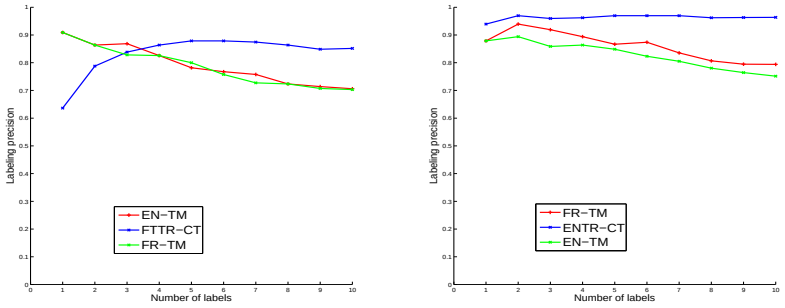

Fig. 4: HANS Evaluation: Comparison of labeling methods with English labels (left) and French labels (right).

3.4 Evaluation on English-French Hansards data

Using HANS, we trained a multilingual topic model for English and French ($L = 2$, $T = 100$ in Sect. 2.2). The number of topics T was chosen by trial and error. A multilingual corpus was constructed as described in Sect. 3.1. We clustered the test set, manually chose coherent clusters, and removed small clusters (size<10). The resulting set of 758 documents in 33 clusters was labeled using four methods (including the gold standard) and in two languages. The results of the evaluation are shown in Fig. 4. It should be interpreted as described in Sect. 3.2.

We also used machine translation[15] to convert the bilingual cluster into a monolingual cluster, and then used the *ctf.cdf.idf* method to do labeling. For English labels (on the left in the figure), this is reported as FRTR-CT (labels generated from French documents (translated into English) using the *ctf.cdf.idf* method). Similarly, we report ENTR-CT for the French labels.

The results suggest that the topic model-based labeling method does reasonably well compared to the gold standard and the machine translation-based method. Also observe that EN-TM and FR-TM almost overlap; i.e. the performance is identical irrespective of the language of the documents. If we consider the fact that it uses no language resources apart from a comparable corpus, this is a very encouraging result.

3.5 Evaluation on English-Hindi-Bengali Wikipedia data

Using WIKI, we trained multilingual topic model for English-Hindi-Bengali ($L = 3$, $T = 50$ in Sect. 2.2). Again, T was chosen by trial and error. We clustered the test set and chose 39 coherent clusters, many of which were small (size<10). To get a higher number of clusters for evaluation, we fixed the minimum cluster size to 4, even though smaller clusters adversely affect the labeling task. The resulting set of 638 documents in 32 clusters was labeled using four methods (including the gold standard) and in three languages. Some example cluster labels and cluster contents are shown in Tables 1 and 2. We observed that labeling

[15] `http://code.google.com/p/google-api-translate-java`

Table 1: WIKI Evaluation: Sample clusters described by labels in English (left), Hindi (center) or Bengali (right).

English labels	**Hindi labels**	**Bengali labels**
philosophy jesus plato kant torah aristotle god hegel nietzsche moses	दर्शन तथा ईश्वर मनुष्य धर्म ईसा भी समाज हम किंतु (philosophy and God man religion Jesus also society we but)	করা তাদের কোন ধর্ম দর্শন দার্শনিক এক তারা সামাজিক ধর্মীয় (do their some religion philosophy philosopher one they social religious)
film films award disney awards hitchcock simpsons chaplin movie academy	चैप्लिन फ़िल्म द जेरी टॉम फिल्म पिट सर्वश्रेष्ठ अभिनेत्री एंड (chaplin film the jerry tom film pitt best actor and)	চলচ্চিত্র পুরস্কার করা ছবি এক দ্য র তাকে সাইরাস ছবির (film prize do film one the r him cyrus film)
ipa dotted languages consonant language unicode vowel colspan dialects alphabet	भाषा पालि शब्द संस्कृत अंग्रेजी साहित्य बोली लिपि स्वर व्याकरण (language pali word sanskrit english literature dialect script vowel grammar)	ভাষা ভাষার করা ভাষায় প্রচলিত প্রায় কথা ইন্দো কোটি উত্তর (language language do language common nearly word indic crore reply)

Table 2: WIKI Evaluation: Sample documents from clusters in Table 1.

Cluster	**Document Titles**
philosophy jesus ...	Aristotle, परिवार (Family), नारीवाद (Feminism), Secularism, Sigmund Freud, দেমোক্রিতোস (Democritus), Anarchism, ईसा मसीह सत्य (Jesus Christ Truth), तत्त्वमीमांसा (Metaphysics), সফিস্ট (Sophist), ...
film films award ...	नाइट राइडर (Knight Rider), ज्योर्ज लुकास (George Lucas), Monica Bellucci, লাদ্রি দি বিচিক্লেত্তে (Ladri di biciclette), Johnny Bravo, Akira Kurosawa, হাওয়ার্ড হুক্স (Howard Hughes), Marilyn Monroe, স্টীবন স্পিলবর্গ (Steven Spielberg), বেনিস ফ়িল্মোত্সব (Venice Film Festival), ...

quality suffered due to (1) Noise (e.g. colspan, র r) (2) Stopwords (e.g. তথা and, করা do) (3) Morphological variants (film-films, ভাষা-ভাষার). In addition to causing label redundancy, morphological variation also caused label *suppression* (because term frequencies are distributed among the variants). Building better stopword lists and stemmers should increase labeling quality significantly.

The results of the evaluation are shown in Fig. 5. It should be interpreted as described in Sect. 3.2. The results are comparable to HANS, except when the labels were generated in Hindi. On closer examination, it was found that the Hindi versions of many (247) documents had very few words (<20). These documents did not significantly influence the gold standard labels in Hindi due to low word frequencies, but they did influence the topic model-based methods (XX-TM) since all topic mixtures have equal weight. On examining the labels, we found that XX-TM labels and the gold labels in Hindi indicated the same topic, but used different words. Hence, in this case, we feel that our choice of gold standard labels might not have been suitable for evaluation.

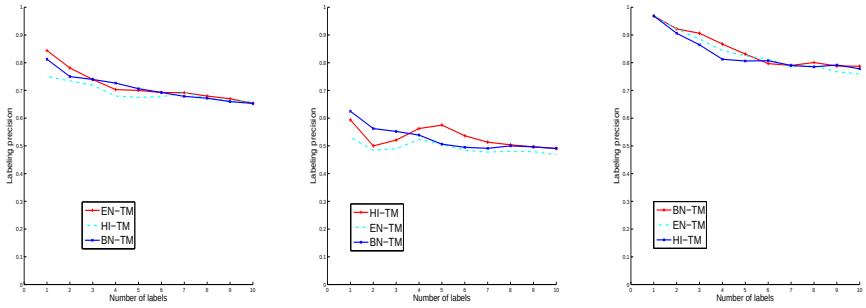

Fig. 5: WIKI Evaluation: Comparison of labeling methods with English labels (left), Hindi labels (center) and Bengali labels (right).

3.6 User Study on English-Hindi-Bengali News

Using the multilingual topic model trained on WIKI (Sect. 3.5), we clustered FIRE and selected 19 coherent clusters and generated labels for them. The clusters were on many topics including Football, Music, Sociology, Movies, etc. We asked two human users to examine each cluster and score its label set as 1 (relevant and coherent), 0.5 (relevant but having noise words), or 0 (unrelated or nonsensical). The score for each label set was the average score over all users; the final score was the average score over all clusters. This is a number between 0 and 1, with higher values indicating better quality. The score, and the agreement metrics, are shown in Table 3. We see that the users found the labels meaningful but noisy. The sources of error were similar to those for WIKI (Sect. 3.5).

The Cohen's κ agreement was substantial for English, and fair for Hindi and Bengali. This did not take into account the ordering of the scores. For example, a disagreement of 0 *vs.* 1 is much worse than 0.5 *vs.* 1. Also, we found that one user (User 2) consistently gave lower scores than the other. Taking these factors into account, we computed a weighted κ [6] and found better agreement. The weight matrix W is given in Table 4. The entry $W_{x,y}$ is the penalty when a label set is given the score x by User 1 and y by User 2. For example, $W_{0.5,0.5} = 0$ since both users gave the same rating (0.5), while $W_{0,1} = 2$ has higher penalty, since it is highly unlikely that User 2 will give a score higher than User 1.

Table 3: FIRE User Study: Average score of label quality (left), and user agreement metrics (right).

Language	**Score**	**Agreemt.**	**Cohen's κ**	**Weighted κ**
English	0.63	79%	0.66	0.87
Hindi	0.58	58%	0.34	0.72
Bengali	0.50	58%	0.36	0.58

Table 4: Weight matrix for weighted κ.

		User 2		
	W	**1**	**0.5**	**0**
User 1	**1**	0	0	1
	0.5	1	0	0
	0	2	1	0

Error Analysis and New Research Problems. Two clusters whose label sets were scored 0 by both assessors were closely examined. One cluster was of financial news. There were very few financial articles in WIKI, and the topic model did not have even one topic on finance. Consequently, it assigned a nearly uniform topic distribution for these documents. The other cluster was about *current* national politics, but the labels indicated that the topic of the cluster was *historical events* of national politics. The political articles in WIKI were mostly of a historical nature. This means that the model correctly identified the topic of the corpus, but it labeled using the (restricted) vocabulary of the training data.

These observations seem to suggest two main sources of erroneous labels: incorrect inference of topic mixtures, and topic word distributions that do not represent the test data. The first error occurs when the domain of the cluster was not seen during training. The second error occurs when the vocabulary of the test cluster is different from the training vocabulary. We feel domain adaptation and *vocabulary* adaptation for multilingual topic models are keys area for future research, and can lead to significant improvements in label quality.

4 Related Work

The first step in most cluster labeling methods is the generation of a list of "important" words from the content of the cluster documents. Several approaches have been tried for this, e.g. using the most frequent terms, frequent phrases [16, 19] or named entities [18] in the cluster, using the top terms in the cluster centroid, and using information gain to identify useful words [8]. However, these methods are most suited for the case when all the documents in the cluster share the same vocabulary, i.e. belong to the same language.

Alternative methods based on linguistic analysis and summarization [17, 4, 12] and using singular value decomposition [10] have also been explored. These methods are either monolingual or require dictionaries/machine translation. Also, they typically aim at generating sentences and not short labels (words/phrases), which are more suitable for our task. In addition, linguistic methods are not portable across languages and hence more expensive.

There have been efforts to improve labeling using external information such as the titles of documents that are close to the cluster centroid [7], anchor text associated with in-links to cluster documents (when they are web pages) [9], WordNet synonyms [5] or Wikipedia concepts and categories [2] related to the "important" terms. The idea is to augment the set of candidate labels, and also to use external information while choosing among the candidates. These methods either assume that the cluster is monolingual or that there already exists a (weighted) list of "important" terms.

5 Conclusion

We have explored multilingual Scatter/Gather in the absence of machine translation, but using comparable corpora. We built a system *ShoBha* based on multilingual topic models and demonstrated effective labeling on several real-world

data sets, and in several languages. We also found that the labels obtained from the topic model were comparable to those obtained using machine translation.

From the user study, we see that domain and vocabulary adaptation of the topic model are important areas for future work. In our experiments, the topic model was also used for multilingual *clustering*. An analysis of this approach has also been left for future work.

Acknowledgments. The authors thank Infosys Limited for supporting this work, and Achintya Kundu and Adway Mitra for their help in the experiments.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I. Latent dirichlet allocation. JMLR (2003)
2. Carmel, D., Roitman, H., Zwerdling, N. Enhancing cluster labeling using wikipedia. SIGIR '09
3. Carpineto, C., Osiski, S., Romano, G., Weiss, D. A survey of web clustering engines. ACM Comput. Surv. (Jul 2009)
4. Chen, H.H., Kuo, J.J., Su, T.C. Clustering and visualization in a multi-lingual multi-document summarization system. ECIR '03
5. Chin, O.S., Kulathuramaiyer, N., Yeo, A.W. Automatic discovery of concepts from text. WI '06
6. Cohen, J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin (October 1968)
7. Cutting, D.R., Karger, D.R., Pedersen, J.O., Tukey, J.W. Scatter/gather: a cluster-based approach to browsing large document collections. SIGIR '92
8. Geraci, F., Pellegrini, M., Maggini, M., Sebastiani, F. Cluster generation and labeling for web snippets: A fast, accurate hierarchical solution. Internet Math. (2007)
9. Glover, E., Pennock, D.M., Lawrence, S., Krovetz, R. Inferring hierarchical descriptions. CIKM '02
10. Honarpisheh, M.A., Ghassem-Sani, G., Mirroshandel, G. A multi-document multilingual automatic summarization system. IJCNLP '09
11. Ke, W., Sugimoto, C.R., Mostafa, J. Dynamicity vs. effectiveness: studying online clustering for scatter/gather. SIGIR '09
12. Kuo, J.J., Chen, H.H. Multidocument summary generation: Using informative and event words. TALIP (February 2008)
13. Manning, C.D., Raghavan, P., Schütze, H. Introduction to Information Retrieval. Cambridge University Press (2008)
14. Mimno, D., Wallach, H.M., Naradowsky, J., Smith, D.A., McCallum, A. Polylingual topic models. EMNLP '09
15. Ming, Z.Y., Wang, K., Chua, T.S. Prototype hierarchy based clustering for the categorization and navigation of web collections. SIGIR '10
16. Osinski, S., Weiss, D. A concept-driven algorithm for clustering search results. IEEE Intell. Sys. (May 2005)
17. Radev, D.R., Jing, H., Styś, M., Tam, D. Centroid-based summarization of multiple documents. Inf. Proc. Manag. (November 2004)
18. Toda, H., Kataoka, R. A clustering method for news articles retrieval system. WWW '05
19. Treeratpituk, P., Callan, J. Automatically labeling hierarchical clusters. Digital Government Research 2006