Mining bilingual topic hierarchies from unaligned text

Sumit Negi
IBM Research India
Vasant Kunj, Institutional Area
New Delhi, India
sumitneg@in.ibm.com

Abstract

Recent years have seen an exponential growth in the amount of multilingual text available on the web. This situation raises the need for novel applications for organizing and accessing multilingual content. Common examples of such applications include Multilingual Topic Tracking, Cross-Language Information retrieval systems etc. Most of these applications rely on the availability of multilingual lexical resources which require significant effort to create. In this paper we present an unsupervised method for building *bilingual topic hierarchies*. In a *bilingual topic hierarchy*, topics (where a topic is a distribution over words) are arranged in a hierarchical fashion with abstract topics appearing near the root of the hierarchy and more concrete topics near the leaves. Such bilingual topic hierarchies can be useful for organizing bilingual corpus based on common topics, cross-lingual information retrieval and cross-lingual text classification. Our method builds upon the prior work done on Bayesian non-parametric inferencing of topic hierarchies and multilingual topic modeling to extract bilingual topic hierarchies from unaligned text. We demonstrate the effectiveness of our algorithm in extracting such topic hierarchies from a collection of bilingual text passages and FAQs.

1 Introduction

The last few decades have seen an explosive growth in Internet accessibility in developing regions of the world. A significant part of this growth has been in countries that use languages other than English as their primary language (Chinese, Spanish, Arabic, Hindi etc). The increasing number of multilingual Internet users has resulted in a tremendous increase in the amount multilingual content that is available on the Web. This situation raises the need for novel ways of organizing and accessing multilingual content. Work done on Cross Language Information Retrieval (CLIR) (Xu et al., 2001) i.e. retrieving information written in a language different from the language of the user's query is one key attempt in this very direction. Similarly, there has been increased interest in multilingual text mining applications such as sentiment detection (Boiy and Moens, 2008), mining multilingual news feeds, cross lingual text categorization (Bel et al., 2003) etc. Most of these applications use bilingual dictionaries or lexical databases to perform such tasks. For instance, (Pouliquen et al., 2004) use the *Eurovoc* multilingual thesaurus for cross-lingual news topic tracking. (Mihalcea et al., 2007) use a bilingual dictionary to generate subjectivity analysis resources for a given language. Similarly, CINDOR (Ruiz et al., 2000) uses interlingua resources to address the cross language information retrieval problem.

Considering the importance of multilingual lexical resources efforts to (semi) automatically build/populate such resources from Web or other large text collections have gained prominence. For example, (Nagata et al., 2001) build a bilingual dictionary of English-Japanese technical term by collecting and scoring translation candidates from the Web. (Widdows et al., 2002; Nerima et al., 2003) describe similar initiatives related to building multilingual lexical resources from large text corpora.

The work presented in this paper is similar in spirit i.e. it presents an automated way of building and populating a lexical resource given a text corpus. In this paper we propose an unsupervised method of building *bilingual topic hierarchies* using Topic models (Blei et al., 2003). In a topic hierarchy, topics (where a topic is a distribution over

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 992–1000,
Chiang Mai, Thailand, November 8 – 13, 2011.

words) are arranged in a hierarchical fashion with abstract topics appearing near the root of the hierarchy and more concrete topics near the leaves. Figure 1 is an example of a bilingual topic hierarchy. Such topic hierarchies can be useful for organizing/navigating bilingual corpus based on common topics, cross-lingual information retrieval and multilingual text categorization.

Motivating Example: Consider a corpus containing medical documents from two different languages on a common set of topics (e.g. genetics, pharmacology, toxicology etc). The documents in this corpus are not aligned in anyway i.e. there is no document or sentence level alignment between documents of the two languages in the corpus. Let us assume that the goal is to organize this bilingual corpus in such a way that it is easy to locate documents of a certain topic/sub-topic irrespective of the language in which the document was authored. One way of achieving this would be to arrange the document from the corpus in some sort of a hierarchy where (a) documents pertaining to different topics appear in different subtrees of the hierarchy and (b) within a given subtree documents belonging to abstract topics appear at higher levels of the subtree than documents belonging to concrete subtopics.

Organizing documents in such a fashion could be aided by the availability of a bilingual topic hierarchy as shown in Figure 1 in which topics (where a topic is a collection of words) are arranged in a hierarchical fashion with abstract topics appearing near the root of the hierarchy and more concrete topics near the leaves. Moreover, using a data-driven approach (as proposed by our method) for building such a bilingual topic hierarchy ensures that the hierarchy is representative of the structure present in the data. Note, that the hierarchy shown in Figure 1 was generated by our proposed method from a corpus containing English and Hindi medical documents on topics such as Behavior (MeSH [1]Category F01.145), Alcohol Drinking (MeSH Category F01.145.317.269) etc.

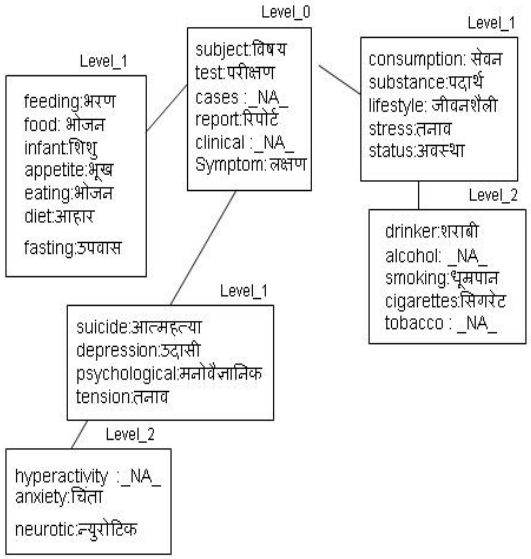

Figure 1: Bilingual Topic Hierarchy

2 Prior Work

Topic models have been used for analyzing topic trends in research literature, inferring captions for images, social network analysis in email, and expanding queries with topically related words in information retrieval. Most of the topic modeling work on text has occurred in the monolingual context. *Multilingual topic models* (MTM) is a relatively new area of research as compared to Topic modeling on monolingual corpus. In their work (Jagarlamudi and Daum III, 2010) (Boyd-Graber and Blei, 2009) demonstrate how a monolingual topic model, which groups together semantically similar words based on similar context, cannot be used directly on a multilingual corpus. This is because of the fact that almost all documents are written in a single language hence similar meaning words from different languages will never share similar context. Consequently, even though a topic model applied on a bilingual corpus will find coherent topics (for each language independently) it will bifurcate the topic space between the two languages.

In Figure 2 we show a few topics discovered by Latent Dirichlet Allocation (Blei et al., 2003) on a bilingual corpus containing Hindi and English documents. The outcome is similar to what was demonstrated by (Jagarlamudi and Daum III, 2010) on the Europarl[2] parallel corpus i.e. LDA bifurcates the topics between the two languages despite their being striking similarity between topics in the two languages. For instance, Topic *C1* and *C3* should be merged together as they form one coherent topic cluster (both these clusters have

[1]Medical Subject Headings

[2]http://www.statmt.org/europarl/

words relevant to *'crop-disease'*). Similarly, Topic *C2* and *C5* should be merged together as they form one coherent topic cluster (both these clusters have words relevant to *'agriculture'*).

In order to discover coherent topics across languages, most multilingual topic models take the document's language into consideration. Previous work on Multilingual Topic Modeling connects the languages by assuming parallelism at either the sentence level or document level. (Zhao and Xing, 2006) propose BiTAM (Bilingual topic admixture models) and HM-BiTAM (Hidden Markov Bilingual topic admixture model), bilingual topic model for Statistical Machine Translation that assume sentence level alignment. Work done by (Kim and Khudanpur, 2004; Tam and Schultz, 2007; Ni et al., 2009) relax the sentence level alignment but require document level alignment. These requirements, namely sentence/document level alignment are too restrictive as finding parallel corpus for all possible domains and languages is difficult. Recent work by (Jagarlamudi and Daum III, 2010) and (Boyd-Graber and Blei, 2009) relax these restrictions by proposing multilingual topic models that work on unaligned text in multiple languages. MuTo (Multilingual Topic Model) (Boyd-Graber and Blei, 2009) does away with the alignment requirement by assuming that similar themes and ideas appear in both languages. The JointLDA model (Jagarlamudi and Daum III, 2010) , which can be seen as a generalization of MuTo, uses a bilingual dictionary to mine bilingual topics from an unaligned corpus. As JointLDA requires only a bilingual dictionary, which is easy to obtain for most pair of languages, we use this model for extracting bilingual topic hierarchies. Please note that the JointLDA model cannot be used directly to infer topic hierarchies. This is because like LDA, JointLDA treats topics as a flat set of probability distributions, with not direct relationship between topics. To discover relationships between topics the Hierarchical Latent Dirichlet Allocation (Blei et al., 2010) is used.

To the best of our knowledge our work of discovering bilingual topic hierarchies using Topic models is a first. Our generative model uses the JointLDA (a multilingual topic modeling approach) and hLDA (a hierarchial topic modeling approach) models to discover bilingual topic hierarchies from an unaligned bilingual corpus. Combining these two models gives us an unsupervised mechanism of discovering bilingual topic hierarchies. Moreover, in this setup no assumption about the topics or hierarchy structure (there are no limitation such as maximum depth or maximum branching factor) needs to be made.

The paper is organized as follows. In Section 3 we provide a short overview of the Hierarchical Latent Dirichlet Allocation and JointLDA. Section 4 provides details of the proposed generative model for extracting bilingual topic hierarchies from an unaligned bilingual corpus. Experiments and conclusion are provided in Section 5 and Section 6 respectively.

3 Hierarchical Latent Dirichlet Allocation and JoinLDA models

For ease of exposition, we first describe the basics of the Hierarchical Latent Dirichlet Allocation and JoinLDA models. Topic models such as Latent Dirichlet Allocation (LDA) treat topics as a flat set of probability distributions, with no direct relationship between topics. While such models can be used to discover set of topics from a corpus they fail to detect the level of abstraction of a topic, or how topics are related. Blei et al. propose a *hierarchical topic model* hLDA (Blei et al., 2010) that learns such relations between topics. Given a collection of documents each of which contains a set of words, hLDA discovers topics in the documents and organizes these topics into a hierarchy. More general topics appear near the root whereas more specialized topics appear near the leaf of the hierarchy. In the hierarchy each node is associated with a topic, where topic is distribution over words. Under this generative model a document is generated by choosing a path from the root to a leaf, repeatedly sampling topics along the path, and sampling the words from the selected topics. Blei et al. define a *nested Chinese restaurant process* which is used as a prior distribution over the possible hierarchies (a hierarchy can be thought of as a infinitely-deep and infinitely-branching tree).

In hLDA each document is assigned a single path in the hierarchy. The first level, which is directly below the root, induces a coarse partition on the documents. Topics at this level place high probability on words that are useful within the corresponding subset/partition. The nested partitions of documents become finer as one moves down the hierarchy. Consequently, the corresponding topics (and the words associated with those topics) be-

C1	C2	C3	C4	C5	C6	C7	C8
pesticide	soil	बीमारी	कीटनाशक	फसल	market	disease	मंडी
pest	harvest	टीका	छिड़कना	गेहूँ	wholesale	vaccine	थोक
blight	grain	परजीवी	बिमारी	चावल	storage	viral	वितरण
rust	cultivate	लक्षण	मिलाना	कपास	district	bacterial	गाँव
parasite	rice	खाल	परजीवी	मौसम	price	symptoms	दाम
	wheat			खेती		influenza	

Figure 2: Topics extracted by LDA from a bilingual Hindi-English corpus

come more specialized to the particular documents in those paths. As mentioned in (Blei et al., 2010) the goal of finding a topic hierarchy at different levels of abstraction is distinct from the problem of hierarchical clustering. Hierarchical clustering treats each data point as a leaf in a tree, and merges similar data points up the tree until all are merged into a root node. Thus, the internal nodes formed during the hierarchical clustering process shares words with their children. In contrast, in the hierarchical topic model the internal nodes are not summaries of their children. Rather, the internal nodes reflect the shared terminology of the documents assigned to the paths that contains them.

We extend the hLDA model to extract *bilingual topic hierarchies* from unaligned bilingual corpus. Even though there has been some attempts to extract topics (not topic hierarchies) from cross-lingual corpus, these approaches assume either explicit or indirect clues about document alignment. In order to overcome such restrictions, namely sentence/document alignment, we use the JointLDA model proposed by (Jagarlamudi and Daum III, 2010). The JointLDA model is an extension of the LDA model which uses bilingual dictionaries to generate documents in different languages. The JointLDA model, like LDA, models a document as a mixture over T topics, where the mixture weight (θ_d) is drawn from a Dirichlet distribution. JointLDA introduces an additional hidden variable called *concepts*, in defining a topic distribution. Each topic is now a distribution over concepts rather than words, where the topic distribution is also drawn from a Dirichlet distribution. Finally, a *concept* can be realized in different ways depending on the choice of the document's language (l_d). This additional layer of language independent abstraction over the words allows the model to capture common topics in different languages effectively. JointLDA use bilingual dictionary entries as substitute for these concepts. Out-of-dictionary words are handled by adding artificial dictionary entries to the dictionary. For more details on the JointLDA model readers should see (Jagarlamudi and Daum III, 2010).

In the next section we describe in detail our proposed bilingual-topic hierarchy generative model.

4 Generative Process

In the bilingual-topic hierarchy model for a given document d a path $\mathbf{c}_d$ (where a path denotes a collection of topics) is drawn from a nested Chinese Restaurant Process (nCRP). Given a choice of a path, or in other words a collection of topics, the GEM (Pitman, 2002) distribution is used to define a probability distribution on the topics along that path. Given a draw from a GEM distribution, a document for language l_d is generated by first repeatedly selecting topics according to the probabilities defined by that draw, followed by selecting a *concept* from a multinomial distribution, and then selecting a word given the *concept* and language of the document. The generative process can thus be summarized as

1. For each table $k \in T$ in the infinite tree
   (a) draw a topic $\beta_k \sim \mathrm{Dir}(\eta)$
2. For each doc $d \in \{1,2....D\}$
   (a) draw $\mathbf{c}_d \sim \mathrm{nCRP}(\gamma)$
   (b) draw a distribution over levels in the tree, $\theta_d|\{\mathrm{m},\pi\} \sim \mathrm{GEM}(\mathrm{m},\pi)$
   (c) for each word in the document
      i. choose level $z_{d,n}|\theta \sim \mathrm{Mult}(\theta_d)$
      ii. select a concept (dictionary entry) $v_{d,n} \sim \mathrm{Mult}(\beta_{\mathbf{c}_d}[z_{d,n}])\Psi(v_{d,n},\mathrm{l}_d)$
      iii. select a word from $\mathrm{p}(w_{d,n}|\ v_{d,n},\mathrm{l}_d)$

Where the function $\Psi(v_{d,n},\mathrm{l}_d)$ is 1 if the dictionary entry $v_{d,n}$ can generate a word from language l_d

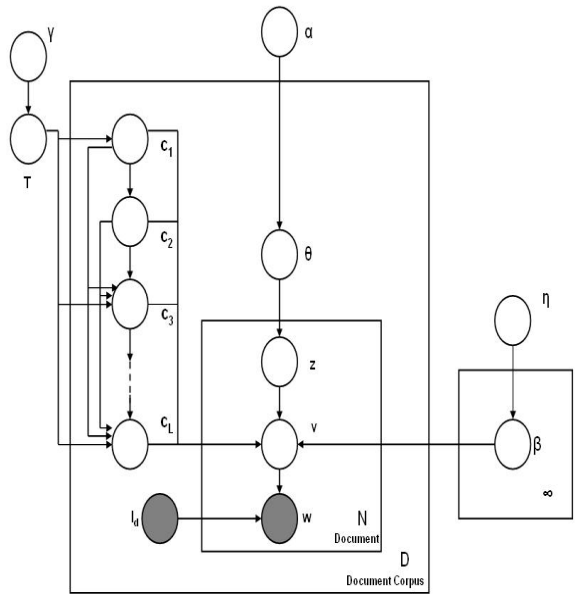

Figure 3: bilingual-topic hierarchy model

and otherwise 0. Given a dictionary entry and language there is only one possibility for a word and hence p($w_{d,n}$| $v_{d,n}$,l_d) = 1 (Jagarlamudi and Daum III, 2010). Figure 3 illustrates the bilingual-topic hierarchy generative process using the plate model notation.

4.1 Approximate Inference

Given the bilingual-topic hierarchy model the goal is to perform *posterior inference* i.e. to invert the generative process of documents for estimating the hidden topical structure of a bilingual document collection. The posterior distribution gives us the distribution of the underlying topic structure that might have generated an observed collection of bilingual documents. Finding this posterior distribution is a key problem in Bayesian statistics. Since the posterior distribution does not have a closed form we employ an approximate inferencing technique namely a variant of Markov Chain Monte Carlo (MCMC) technique called *collapsed Gibbs sampling* (Liu, 1994). In a Gibbs sampler each latent variable is iteratively sampled conditioned on the observations and all other latent variables. To speed up convergence the collapsed Gibbs sampler marginalize out some of the latent variables. Collapsed Gibbs sampling for topic models has been widely used in topic modeling applications (McCallum et al., 2004; Mimno and McCallum, 2007; Rosen-Zvi et al., 2004).

The variables needed by the sampling algorithm are: $w_{d,n}$: the *n*th word in document *d* (observed), l_d : document *d*'s language (observed), $c_{d,l}$: the *l*th topic in document *d*, $v_{d,n}$: the *concept* (dictionary entry) associated with the *n*th word in document *d* and $z_{d,n}$: the assignment of the *n*th word in document *d* to one of the L available topics. The Gibbs sampler integrates out all the other variables in the model to assess the values of $z_{d,n}$, $c_{d,l}$ and $v_{d,n}$.

We approximate the posterior
p($\mathbf{c}_{1:D}$, $\mathbf{z}_{1:D}$, $\mathbf{v}_{1:D}$ | η, γ, m, π, $\mathbf{w}_{1:D}$, $\mathbf{l}_{1:D}$)
where hyper-parameter γ reflects the likelihood that documents will choose new paths when traversing the nested CRP, η reflects the expected variance of the underlying topics ($\eta \ll 1$ will tend to choose topics with fewer high probability words), and m and π reflect our expectation about the allocation of words to levels within a document. Bold fonts $\mathbf{c}_{1:D}$, $\mathbf{z}_{1:D}$, $\mathbf{v}_{1:D}$ denote vector of level allocations, topic allocations and concept allocation respectively. The variable $\mathbf{c}_d$ denotes the per-document path, $z_{d,n}$ denotes per-word level allocation to topics in those paths (as mentioned in Section 3, a path in a tree picks out a collection of topics) and $v_{d,n}$ denotes *concept* for the *n*th word in document *d*. Next, we describe in detail how level allocations, concepts and paths are sampled.

4.1.1 Sampling

In this section Equation (1), (2), (3) detail how the level allocation variable $z_{d,n}$ is sampled. Given the current path assignments, we sample the level allocation variable $z_{d,n}$ for word *n* in document *d* from its distribution given the current value of all other variables

$$p(z_{d,n}|\mathbf{z}_{-(d,n)}, \mathbf{c}, \mathbf{v}, m, \pi, \eta) \propto p(z_{d,n}|\mathbf{z}_{d,-n}, m, \pi).p(v_{d,n}|\mathbf{z}, \mathbf{c}, \mathbf{v}_{-(d,n)}, \eta) \quad (1)$$

Where $\mathbf{z}_{-(d,n)}$ and $\mathbf{v}_{-(d,n)}$ are vectors of level allocations and concepts leaving out $z_{d,n}$ and $v_{d,n}$ respectively. The first term in (1) is a distribution over levels.

$$p(z_{d,n} = k \mid \mathbf{z}_{d,-n}, m, \pi) = \frac{(1-m)\pi + \#[\mathbf{z}_{-(d,n)} = k]}{\pi + \#[\mathbf{z}_{-(d,n)} \geq k]} \prod_{j=1}^{k-1} \frac{m\pi + \#[\mathbf{z}_{d,-n} > j]}{\pi + \#[\mathbf{z}_{d,-n} \geq j]} \quad (2)$$

Where #[..] counts the elements of an array satisfying a given condition. Interested readers are requested to refer to (Blei et al., 2010) for detailed

derivation. The second term in (1) is the probability of a given concept based on possible assignment. From the assumption that the topic parameter β_i are generated from a dirchilet distribution with hyper-parameter η we obtain

$$p(v_{d,n} = j \mid \mathbf{z}, \mathbf{c}, \mathbf{v}_{-(d,n)}, \eta) \propto \#[\mathbf{z}_{-(d,n)} = z_{d,n}, \mathbf{c}_{z_{d,n}} = c_{(d,z_{d,n})}, \mathbf{v}_{-(d,n)} = j] + \eta \quad (3)$$

Where (3) gives the smoothed frequency of the number of times concept *j* is allocated to topic at level $z_{d,n}$ of the path c_d.

Given level allocation, in order to sample the path associated with each document conditioned on all other paths and concept assignments (4), (5) is used. Where

$$p(\mathbf{c}_d \mid \mathbf{c}_{-d}, \mathbf{v}, \mathbf{z}, \eta, \gamma) \propto p(\mathbf{c}_d \mid \mathbf{c}_{-d}, \gamma).p(\mathbf{v}_d \mid \mathbf{c}, \mathbf{z}, \mathbf{v}_{-d}, \eta). \quad (4)$$

The probability of *concept* is obtained by integrating over the multinomial parameters.

$$p(\mathbf{v}_d \mid \mathbf{c}, \mathbf{z}, \mathbf{v}_{-d}, \eta) = \prod_{l=1}^{max(z_d)} \frac{\Gamma(\sum_v \#[\mathbf{z}_{-d} = l, \mathbf{c}_{-d,l} = c_{d,l}, \mathbf{v}_{-d} = v] + V\eta)}{\prod_v \Gamma(\#[\mathbf{z}_{-d} = l, \mathbf{c}_{-d,l} = c_{d,l}, \mathbf{v}_{-d} = v] + \eta)} \times \frac{\prod_v \Gamma(\#[\mathbf{z} = l, \mathbf{c}_l = c_{d,l}, \mathbf{v} = v] + \eta)}{\Gamma(\Sigma_v \#[\mathbf{z} = l, \mathbf{c}_l = c_{d,l}, \mathbf{v} = v] + V\eta)} \quad (5)$$

For each document *d* the topic and concept assignments for each word i.e. ($z_{d,n}$,$v_{d,n}$) is sampled from the probability distribution given by

$$p(z_{d,n} = k, v_{d,n} = j \mid \mathbf{z}_{-(d,n)}, \mathbf{v}_{-(d,n)}, \mathbf{w}, \mathbf{l}) \propto \frac{n^d_{-(d,n),k} + \gamma}{n^d_{-(d,n),(.)} + L\gamma} \cdot \frac{n^j_{-(d,n),k} + \eta}{n^{(.)}_{-(d,n),k} + V\eta} \quad (6)$$

In (6) $n^j_{-(d,n),k}$ ($n^{(.)}_{-(d,n),k}$) is the number of times the dictionary entry *j* (any dictionary entry) is used along with topic *k* for sampling any word excluding the word $w_{d,n}$. Similarly, $n^d_{-(d,n),k}$ ($n^d_{-(d,n),(.)}$) is the number of words in document *d* that are assigned to topic *k* (any topic) excluding the word $w_{d,n}$. Note, L is the number of topics (levels in the hierarchy) and V the vocabulary size. Figure 4 provides an overview of the sampling algorithm.

Given the current state of the sampler [$\mathbf{c}_{1:D}$, $\mathbf{z}_{1:D}$, $\mathbf{v}_{1:D}$]
Begin
  For each document *d* ∈ {1,.....,D}
    Draw $\mathbf{c}_d$ using (4)
    For each word in document *d* draw $z_{d,n}$ using (1)
    For each word in document *d* draw $v_{d,n}$ using (6)
End

Figure 4: Gibbs Sampling Algorithm

5 Experiments

To demonstrate that the bilingual-topic hierarchy model extracts meaningful topic hierarchies the model was applied on two real world data-sets. The first data-set is a collection of 2000 questions on the following three topics: *health insurance*, *auto insurance* and *passport/visa* queries. This data set was built by crawling government and insurance company web sites. The average length of a question in this corpus is eleven words. The question corpus contains 1200 questions in Hindi and 800 question in English. Further details of this data-set, henceforth referred to as *Data-Set A*, is provided in Table 1. The extracted bilingual topic hierarchy is shown in Figure 5. For our experiments we used *Shabdanjali*[1] as our bilingual (Hindi-English) dictionary. Since the coverage of a bilingual dictionary will be limited we add artificial entries ($\_NA\_$) for out-of-dictionary words. This is similar to the approach adopted by (Jagarlamudi and Daum III, 2010) in their JointLDA work.

Language	Health	Auto	Passport/Visa	Total
Hindi	440	330	430	1200
English	280	350	170	800

Table 1: Data-Set A.

We apply our algorithm on a second data set which is a collection of text passages on *agricultural* and *animal rearing*. This data-set, henceforth referred to as *Data-Set B*, is a collection of 1100 passages, 700 of which are in Hindi and the rest 400 in English. The average length of a passage in this data-set is 221 words. The extracted bilingual topic hierarchy is shown in Figure 6.

In order to facilitate the visualization of extracted topic hierarchies we restrict the number of levels to three. As is evident from Figure 5, Figure 6 the model was able to successfully extract the underlying bilingual-hierarchical struc-

[1] http://www.shabdkosh.com/content/category/downloads/

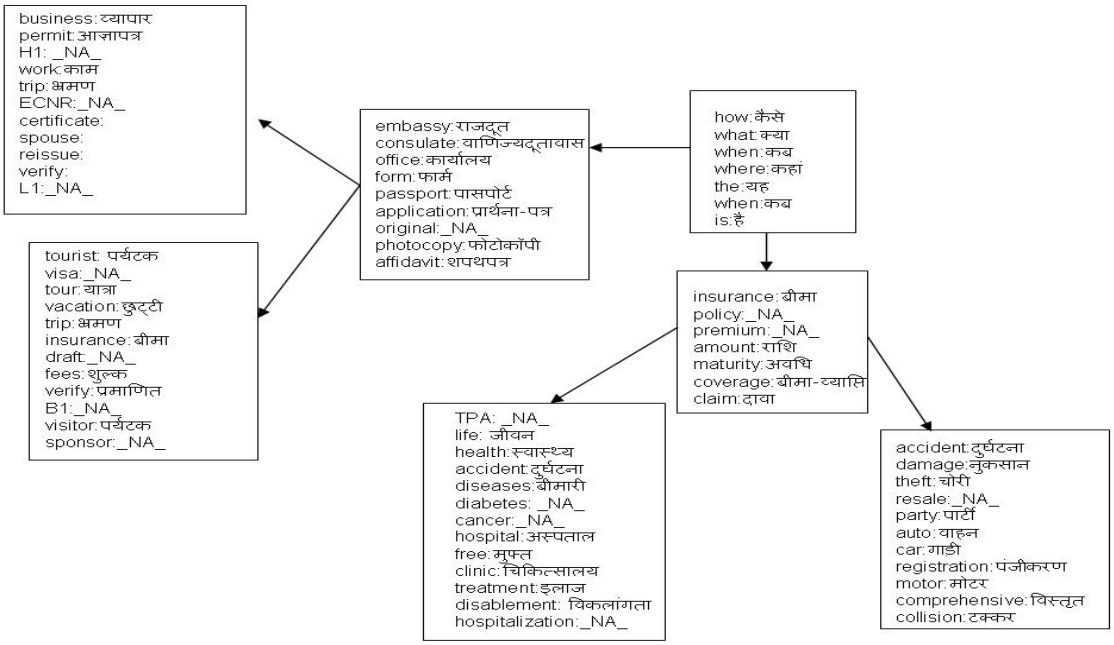

Figure 5: Bilingual Topic Hierarchy: Data-Set A

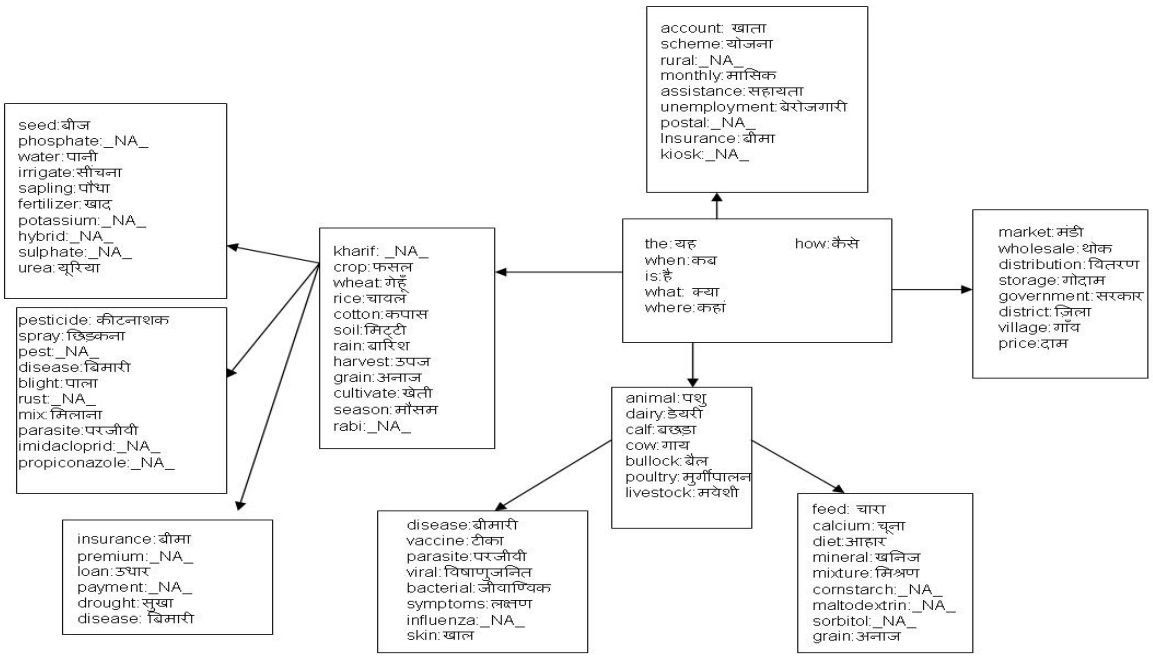

Figure 6: Bilingual Topic Hierarchy: Data-Set B

ture from the two document collections. For example for *Data-Set B* (Figure 6) the second level of the hierarchy captures the two prominent topics (where a topic is a distribution over words) present in the document collection namely that related to agriculture and animal rearing. The third level of this hierarchy further refines these topics. The *agriculture* topic is further split into subtopics related to crop-disease, crop-cultivation and crop-insurance. Similarly, the *animal-rearing* topic cluster is split into subtopic related to animal-disease and animal-feed. For a quantitative evaluation of our method we use predictive held-out likelihood as a measure of performance. Fig-

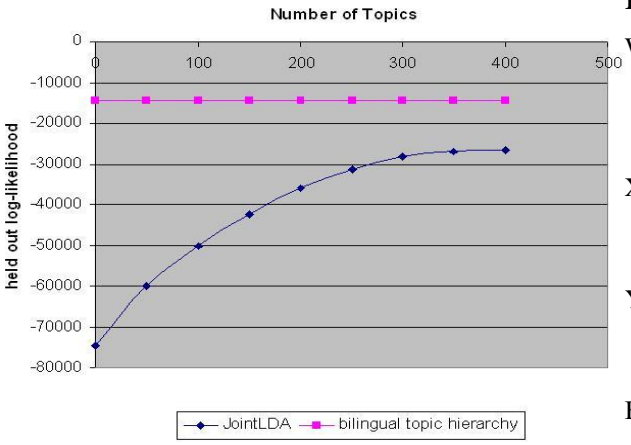

Figure 7: Held out log-likelihood vs Number of topics

ure 7 illustrates the five-fold cross-validated held-out likelihood for JointLDA and bilingual topic hierarchy on the test corpus. It is evident from this experiment that for topic cardinalities in the range 0 to 400, bilingual topic hierarchy provides significantly better predictive performance than JointLDA.

6 Conclusion

In this paper we presented an unsupervised method for building *bilingual topic hierarchies*. In a topic hierarchy, topics (where a topic is a distribution over words) are arranged in a hierarchical fashion with abstract topics appearing near the root of the hierarchy and more concrete topics near the leaves. Such bilingual topic hierarchies can be useful for organizing bilingual corpus based on common topics, cross-lingual information retrieval and cross-lingual text classification. We propose a generative model that employs Bayesian non-parametric inferencing of topic hierarchies and multilingual topic modeling to extract such bilingual topic hierarchies from unaligned text. The effectiveness of the algorithm in extracting bilingual topic hierarchies is demonstrated on a collection of bilingual text passages and FAQs. As part of future work we plan to use the extracted bilingual topic hierarchies for cross-lingual text classification and information retrieval tasks.

References

W. Kim and S. Khudanpur. 2004. *Lexical triggers and latent semantic analysis for cross-lingual language model.* ACM Transactions on Asian Language Information Processing.

X. Ni and J.T Sun and J. Hu and Z. Chen. 2009. *Mining multilingual topics from Wikipedia.* In International World Wide Web Conference.

Y. C Tam and T. Schultz. 2007. *Bilingual LSA-based translation lexicon adaptation for spoke language translation.* In INTERSPEECH.

B. Zhao and E.P Xing. 2006. *Bilingual topic admixture models for word alignment.* Association for Computational Linguistics.

B. Pouliquen and R. Steinberger and C. Ignat and E. Kasper and I. Temnikova. 2004. *Multilingual and Crosslingual News Topic Tracking.* In COLING 2004.

J. Xu and R. Weischedel and R. Nguyen. 2001. *Evaluating a probablistic model for cross-lingual information retrieval.* In SIGIR 2001.

N. Bel and C.H.A Koster and M. Villegas. 2003. *Cross-lingual text categorization.* In ECDL 2003.

E. Boiy and M. Moens. 2008. *A machine learning approach to sentiment analysis in multilingual Web texts.* Information Retrieval, 2008.

M. Nagata and T. Saito and K. Suzuki. 2001. *Using the web as a bilingual dictionary.* In Proceedings of the workshop on Data-driven methods in machine translation.

D. Widdows and B. Dorow and C. Chan. 2002. *Using parallel corpora to enrich multilingual lexical resources.* In Third International Conference on Language Resources and Evaluation.

L. Nerima and V. Seretan and E. Wehrli. 2003. *Creating a multilingual collocation dictionary from large text corpora.* In Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics.

J. Bernardo and A. Smith. 1994. *Bayesian Theory.* John Wiley & Sons Ltd.

D. Blei and T. Griffiths and M. Jordan. 2010. *The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies.* Journal of the ACM.

J. Jagarlamudi and H. Daum. 2010. *Extracting Multilingual Topics from Unaligned Comparable Corpora.* 32nd European Conference on IR Research,ECIR.

J. Boyd-Graber and D. M. Blei. 2009. *Multilingual Topic Models for Unaligned Text.* Uncertainty in Artificial Intelligence.

J. Pitman. 2002. *Combinatorial Stochastic Processes.* Lecture Notes for St. Flour Summer School.

J. Liu. 1994. *The collapsed Gibbs sampler in Bayesian computations with application to a gene regulation problem.* Journal of the American Statistical Association.

A. McCallum and A. Corrada-Emmanuel and X. Wang. 2004. *The author-recipient-topic model for topic and role discovery in social networks: Experiments with Enron and academic email.* Tech. report, University of Massachusetts, Amherst.

D. Mimno and A. McCallum. 2007. *Organizing the OCA: Learning faceted subjects from a library of digital books.* In Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital libraries.

M. Rosen-Zvi and T. Griffiths and M. Steyvers and P. Smith. 2004. *The author-topic model for authors and documents.* In Proceedings of the 20th Conference on Uncertainity in Artificial Intelligence.

M. Ruiz and A. Diekema and P. Sheridan. 2000. *CINDOR Conceptual Interlingua Document Retrieval: TREC-8 Evaluation.* In Proceedings of the Eighth Text Retrieval Conference (TREC8).

R. Mihalcea and C. Banea and J. Wiebe. 2007. *Learning multilingual subjective language via cross-lingual projections.* In Proceedings of the Association for Computational Linguistics (ACL).

D. M. Blei and A. Y. Ng and M. I. Jordan. 2003. *Latent dirichlet allocation.* Journal of Machine Learning Research.