Labeling Unlabeled Data using Cross-Language Guided Clustering

Sachindra Joshi
IBM Research
New Delhi, India
jsachind@in.ibm.com

Danish Contractor
IBM Research
New Delhi, India
dcontrac@in.ibm.com

Sumit Negi
IBM Research
New Delhi, India
sumitneg@in.ibm.com

Abstract

The effort required to build a classifier for a task in a *target* language can be significantly reduced by utilizing the knowledge gained during an earlier effort of model building in a *source* language for a similar task. In this paper, we investigate whether unlabeled data in the target language can be labeled given the availability of labeled data for a similar domain in the source language. We view the problem of labeling unlabeled documents in the target language as that of clustering them such that the resulting partitioning has the *best* alignment with the classes provided in the source language. We develop a cross language guided clustering (CLGC) method to achieve this. We also propose a method to discover concept mapping between languages which is utilized by CLGC to transfer supervision across languages. Our experimental results show significant gains in the accuracy of labeling documents over the baseline methods.

1 Introduction

The last few years have seen a rapid growth in the development of machine learning applications for non-English languages. This growth can be attributed to several factors such as increased Internet penetration (especially in non-English speaking countries) and wide adoption of Unicode standards that allow people to generate content in their own language.

A key guiding principal in the development of such applications for a new language (referred to as the target or resource-poor language) has been to leverage the existing models and linguistic resources available for a popular language such as English (also called source or resource-rich language). Existing literature examines two ways of utilizing this knowledge. The first way is to adapt an existing statistical model for a new target language. Examples of this is the problem of cross-lingual sentiment classification (Xiaojun Wan 2009), or in a more general setting for cross language domain adaptation for classification (Peter Prettenhofer and Benno Stein 2010). The second way is to develop linguistic resources for a target or resource-poor language by leveraging the resources available in a source or resource-rich language. An example of this is the work done for automatically transferring syntactic relations (in WordNet) from a source language (English) into a target language (Romanian) (Verginica Barbu Mititelu and Radu Ion 2005).

In this paper, we investigate another way of utilizing the knowledge gained in one language for building machine learning applications in an another language. Our work focuses on generating training data (in contrast to adapting models and language resources) in the target language, given in-domain training data for the source language. The labeled data in the source language could be used to guide the grouping of unlabeled data in the target language, where each group *aligns* to a class label from the source language. We assume that the domain for both the source and target language data is similar and therefore the set of class labels across the two languages will be shared (but may not be exactly the same). As an example consider a real world scenario from a call routing application. A call routing application maps natural language utterances (typically a caller's response to an open ended question such as "how may I help you") to one of a given set of classes also called call types. Figure 1 shows examples of a few ut-

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 383–391,
Chiang Mai, Thailand, November 8 – 13, 2011.

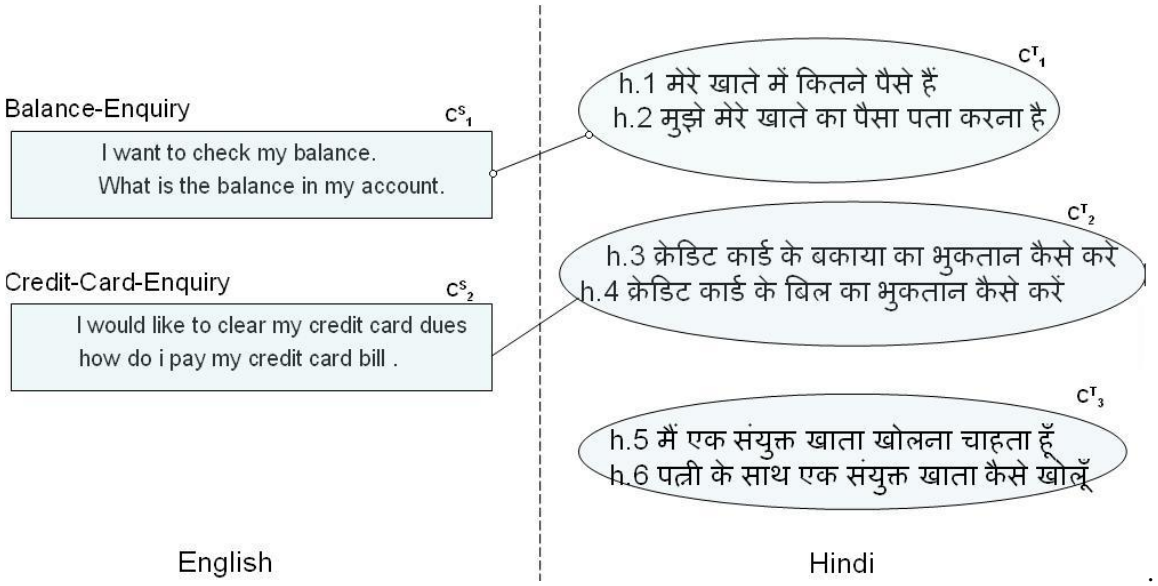

Figure 1: Utterances and class labels in source and target languages

terances (in English) along with associated class labels from the banking domain. These labeled utterances could be used as training data for building a call-routing classifier for the two class labels namely "Balance-Enquiry" and "Credit-Card-Enquiry". Let us assume that we now have utterances in a new language (in Hindi) which are unlabeled. Given that these utterances belong to the same domain, they can be labeled using the same label set as the one used for the source language. This is shown in the Figure 1 where utterances $h.1$ and $h.2$ are grouped together and labeled as "Balance-Enquiry" and utterance $h.3$ and $h.4$ is labeled as "Credit-Card-Enquiry". The labeled data can then be used to train a classifier in the target language.

To label the target language documents automatically we propose a method called cross-language guided clustering (CLGC). This method is built upon a recently proposed approach called cross guided clustering (CGC). CGC guides clustering of documents in a target domain given clusters/classes in a source domain (Bhattacharya et al. 2009). This is achieved by discovering a partitioning in the target domain that is most "similar" or "aligned" to a given partitioning in the source domain. In CLGC we view the problem of labeling unlabeled documents in the target language as that of clustering them such that the resulting partitioning has the *best* alignment with the classes provided in the source language. Since in our case the source and target data are in different languages, we extend the CGC framework to transfer supervision across different languages. We develop cross language similarity measures that use word level and concept level mappings to guide the clustering across languages. We also develop methods to discover concept level mapping between languages. Our experimental results show significant gains in the accuracy of labeling documents over the baseline methods.

One could argue that if the final goal is to classify documents in the target language, this could be achieved by either of the following approaches - (1) by adapting the source language classifier (Peter Prettenhofer and Benno Stein 2010) or (2) by translating unlabeled documents from the target language to the source language and then applying a source language classifier (Mckeown et al. 2003). We claim that our approach is more general and has several advantages over both these approaches. First, building a classifier given a training dataset is a well studied and understood problem. Several off-the-shelf machine learning tools exist that can readily be used for tasks such as feature construction, and building classifiers, provided a training dataset is available (Hall et al. 2009). Our approach can be used to generate a training dataset for the target language which enables use of existing approaches not only for building classifiers, but also for feature engineering tasks such as feature construction and feature selection. This cannot be done using either of the above mentioned approaches.

Second, a key assumption made in both these approaches is that the class labels across languages

are completely shared. This may not be true in several cases as there could be categories that are specific to the target language dataset. As an example, while most of the Hindi utterances in the Figure 1 can be grouped and aligned with a class label in the source language, there exist utterances ($h.5$,$h.6$) which do not belong to any of the existing labels in the source language. Our method allows such groupings to be discovered which can then be used to build target language specific class labels. Moreover, it is worth mentioning that apart from these advantages our proposed method is more efficient than machine translation based methods as it does not require a complete machine translation system.

The specific contributions made by us in this paper are two fold. First, we introduce the problem of labeling documents in one language using the set of labeled documents in another language and show that it is not only feasible but also better than other competitor techniques. Second, we extend the CGC framework to transfer supervision across languages. For this we develop methods to discover concept level mapping between languages that is utilized to guide the clustering across languages.

The rest of the paper is organized as follows. In Section 2 we present related work. We formulate the problem in Section 3. We describe the cross-language guided clustering framework in Section 4. In Section 5, we describe the cross language similarity measure that is used in the CLGC framework. We provide the experimental results in Section 6 and conclude in Section 7.

2 Prior Work

The two research areas that are related to our work are, (1) cross lingual classification and clustering, and (2) semi-supervised clustering.

Cross Lingual Classification and Clustering : Traditional approaches to cross language text classification use linguistic resources such as bilingual dictionaries or parallel corpora to induce correspondences between two languages (Olsson 2005). Some of these methods employ latent semantic analysis (LSA) (Dumais et.al. 1997) or kernel canonical correlation analysis, CCA (Fortuna and Shawe-Taylor 2005). The major limitations of these approaches are their computational complexity and dependence on a parallel corpus. Cross-lingual clustering aims to cluster a heterogeneous (a collection of documents from different languages) document collection. Initial work done in cross-lingual document clustering employed an expensive machine translation (MT) system to fill the gap between two languages (Mckeown et al. 2003). Later work (Wu 2007) done in this area demonstrated that it was possible to achieve comparable performance to the direct MT method using simple linguistic resource such as bilingual dictionaries.

Semi-supervised clustering: Semi-supervised clustering aims to improve clustering performance by limited supervision in the form of a small set of labeled instances. Alternatively, a small set of labeled instances can be used to learn a parameterized distance function (M. Bilenko and R. J. Mooney 2003), (Klein et al. 2002). The co-clustering approach (Dhillon et al. 2003), (N. Slonim and N. Tishby 2000) clusters related dimensions simultaneously through explicitly provided relations between them, such as words and documents, or people and reviews.

The problem that we address in this paper differs significantly from the above mentioned work. Unlike others, our objective is to cluster target language documents such that the resulting clusters are most *'similar'* or best *'aligned'* to the given source language classes. This problem is an instance of semi-supervised clustering in a bilingual setting, which to the best of our best knowledge has received very little attention. Our work builds upon Cross Guided Clustering (CGC) work (Bhattacharya et al. 2009) where supervision is discovered in the form of cluster level similarities obtained from labeled instances from a different domain, having different but related labels. In our work we extend the CGC framework to transfer supervision across different languages.

3 Problem Formulation

Let $T^S = \{< d_1^S, l_1^S >, < d_2^S, l_2^S >, \ldots, < d_n^S, l_n^S >\}$ denote a training dataset in the source language S for a classification task γ. Here $d_i^S \in D^S$ denotes a document that has an associated class label $l_i^S \in L^S$ where, L^S denotes the set of class labels used in T^S. Note, that L^S induces a partitioning of D^S, where each class label l_i^S can be seen as a cluster containing documents d_i^S that have l_i^S as the class label. We are also given a set of unlabeled documents $D^T = \{d_1^T, d_2^T, \ldots, d_m^T\}$ where all the documents are from a similar domain as in T^S but are from a different language T. Our objective is to generate a training dataset using D^T

for the classification task γ. We pose this as a clustering problem over document set D^T, where the resulting clusters are *aligned* with the given classes in the source language dataset. The alignment is achieved by taking the supervision from the partitioning of D^S, which is induced by the label set L^S, to guide the clustering of document set D^T. We refer to this clustering method as *cross-language guided clustering*. In the next section, we describe cross-language guided clustering in detail.

4 Cross-Language Guided Clustering

In this section, we modify the cross guided clustering framework as described in (Bhattacharya et al. 2009) to transfer supervision across languages. Let $Dis(d_i^T, d_j^T)$ provide a distance measure between documents d_i^T and d_j^T in the target language T. A clustering method partitions the given document set into k clusters denoted by centroids $C^T = \{C_1^T, C_2^T, \ldots, C_k^T\}$ such that the total divergence $Div(C^T)$ also referred to as *target only divergence* is minimized. This is defined as follows.

$$Div^T(C^T) = \sum_{C_i^T} \sum_{d_j^T} \delta(C_i^T, d_j^T) Dis(C_i^T, d_j^T)^2 \quad (1)$$

Here $\delta(C_i^T, d_j^T)$ returns 1 if d_j^T is assigned to the centroid C_i^T else returns 0. This is a standard formulation used in the K-Means algorithm (Hall et al. 2009).

In our problem setting, we are additionally provided with a labeled dataset in the source language where the label set induces a partitioning $C^S = \{C_1^S, C_2^S, \ldots, C_l^S\}$ of D^S in the source language. Our objective is to discover partitioning of D^T such that each resulting cluster is aligned with *at most* one class label from the source language and vice-versa. This enables discovery of clusters in the target language that are aligned with the classes in the source language while simultaneously allowing for discovery of any additional concept in the target language. To do this, we require a cross-language similarity function $Sim^X(..)$ that given two documents from different languages, returns a similarity score. This is non-trivial as documents in different languages are represented in entirely separate attribute/feature space. We develop a cross-language similarity measure to achieve this in Section 5. For now, we assume that we have access to such a measure.

To find a cross-language alignment between the source partition and the target partition we construct a bipartite cross language graph G_x that has one set of vertices C^S corresponding to source centroids, and another set C^T corresponding to target centroids. An edge is added between every pair of vertices (C_i^S, C_j^T) where the weight of the edge is given by $Sim^X(C_i^S, C_j^T)$. Now finding the best cross language alignment is equivalent to finding the maximum weighted bipartite match in the graph G_x. Recall that a matching is a subset of the edges such that any vertex is spanned by at most one edge. The score of a matching is the sum of the weights of all the edges in it. In our implementation, we use the 'Hungarian method' to determine the matching (Kuhn 1955).

The matching provides an alignment between the source classes and the target clusters. We only consider those edges in the matching whose weight is more than some predefined threshold. To measure the goodness of cross-language alignment we define a cross-language divergence measure:

$$Div^X(C^S, C^T) = \sum_{C_i^S} \sum_{C_j^T} \delta^X(C_i^S, C_j^T)(1 - Sim^X(C_i^S, C_j^T))^2 |C_j^T| \quad (2)$$

Here, $\delta^X(C_i^S, C_j^T)$ returns the weight of the edge between node C_i^S and node C_j^T if these nodes are matched, else it returns 0. Here $|C_j^T|$ denotes the size of the cluster for which C_j^T is the centroid. The weighing by $|C_j^T|$ is done to make $Div^X(C^S, C^T)$ comparable to $Div(C^T)$. Now the combined divergence between the source partition and the target partition is computed by taking a weighted sum of target-only divergence and cross-language divergence.

$$Div(C^S, C^T) = \alpha * Div^T(C^T) + (1 - \alpha) * Div^X(C^S, C^T) \quad (3)$$

Here α captures the relative importance of the two divergences.

We now provide an algorithm (see Figure 2) that minimizes the objective function given in Equation 3. The algorithm starts by selecting k random data points as centroids from the target language and then executes the following two steps in each iteration. It first assigns points to their nearest centroids and then re-estimates the target centroids to minimize cross-language divergence as given in Equation 3. This is achieved by the

Procedure CrossLanguageGuidedClustering

Select k centroids randomly from D^T
% Initialize target clusters
Iterate n times or until convergence
  Iterate m times
    Assign each $d_i^T \in D^T$ to the nearest centroid
    Recompute the centroids

  % Start CLGC
  Create cross language similarity graph G_x using Sim^X
  Compute maximum bipartite graph matching over G_x
  Iterate over k target centroids in C^T
    Update centroid using the cross language update rule
    Assign each $d_i^T \in D^T$ to the nearest centroid
Return k centroids

Figure 2: Procedure for Cross Language Guided Clustering

following update rule that is obtained by differentiating the divergence function in Equation 3 with respect to the current target centroids.

$$C_i^T = \frac{\alpha \sum_{d_i^T \in C_i^T} d_i^T + (1-\alpha) \sum_j \delta^x(C_i^T, C_j^S)\phi(C_j^S)}{\alpha|C_i^T| + (1-\alpha)|C_i^T| \sum_j \delta^x(C_i^T, C_j^S)\phi(C_j^S)} \quad (4)$$

Here the δ^X function captures the current matching of target clusters with source classes. Intuitively, there are two factors contributing to the update rule. The first factor tries to move the current target centroid towards the center of the cluster computed using the currently assigned data points. This is similar to the standard K-means approach. The second factor that arises due to cross-language alignment tries to move the centroid towards the currently matched source class. Since the feature space used to represent source classes and target centroids are different, we use the function ϕ that projects source classes in the feature space used by the target language. We provide more details regarding the projection function and cross-language similarity in the next section.

5 Cross Language Similarity

In order to perform cross language guided clustering we need a similarity function Sim^X that given two documents d_i^S and d_j^T from source and target languages, computes a similarity score. Let V^S and V^T be the vocabularies used to represent documents in source and target language respectively. Given a word $w_i^S \in V^S$, let the function $proj(w_i^S)$ return a probability distribution $P = \{p_1, p_2, \ldots, p_{|V^T|}\}$ where p_j represents the probability of the word w_i^S being translated to the word w_j^T in target dictionary. The function $proj(..)$ has access to a statistical dictionary $\mathcal{D}_T^S$ for doing this. The dictionary could be constructed using some large general purpose parallel corpus. We now present three different methods to compute the similarity function $Sim^X(d_i^S, d_j^T)$.

Projection based Method: Let M represent a matrix of dimension $|V^S| * |V^T|$ where each i^{th} row contains the probability distribution returned by $proj(w_i^S)$ for $1 \leq i \leq |V^S|$. Given a source document d_i^S, let $\bar{d}_i^S$ refer to its vector representation using the feature space V^S. Then the projection function $\phi(\bar{d}_i^S) = (\bar{d}_i^S)' M$ and the similarity function Sim^X can be defined as follows, where $'$ denotes transpose of a matrix:

$$Sim^X(d_i^S, d_j^T) = \phi(\bar{d}_i^S)\bar{d}_j^T = (\bar{d}_i^S)' M \bar{d}_j^T \quad (5)$$

Weighted Projection based Method: The function $proj(w_i^S)$ returns a probability distribution that captures the likelihood that w_i^S gets translated to a word w_j^T in the target dictionary. Since, this function uses a general purpose bi-lingual statistical dictionary it does not capture domain specific translations. For example, the English word "bank" may have equal probabilities for being translated as "बैंक" or "किनारा" however, given a corpus from the banking domain, it is more likely that the word "bank" translates to "बैंक". Therefore, given a source term we weigh the probability values of the target terms that it translates to, by the frequency of the target terms computed over the target corpus. We then normalize these values again to obtain a probability distribution.

Semantic Mapping based Method: There are multiple words that are synonymous to each other and can be used to represent the same meaning. For example, the word "games" and "sports" are synonymous English words and can be used to represent the same meaning as "खेल" or "गेम्स". The matrix M used in the previous methods, captures the translation probabilities at the word level. In this method we first discover the concepts in each language and then find translation probabilities at the concept level. We refer to this as semantic mapping between the two languages.

To discover the concepts, words from the source and target vocabulary are clustered into *term clusters* based on the words that occur in its *context*. For this a word-by-word co-occurrence matrix is built for the given language. The entry (i, j) in the matrix contains the number of times the word w_i and w_j occur within a fixed window of L

words in the corpus. Thus, each word is represented by a vector called "context vector" that captures words occurring in the context of the given word. We then use an off-the-shelf clustering algorithm (Hall et al. 2009) to obtain term clusters in a language. These term clusters are referred to as concepts. The Figure 3 shows examples of concepts identified in English and Hindi languages. Let $G^S = \{G_1^S, G_2^S, \ldots, G_l^S\}$ and $G^T = \{G_1^T, G_2^T, \ldots, G_m^T\}$ be the source and target concepts obtained by clustering. To find the semantic relationship across concepts from different languages, we construct a bipartite graph that has one set of vertices G^S corresponding to the source concepts, and another set G^T corresponding to the target concepts. Now for each word $w^S \in G_i^S$, we determine the set of target words T_{w^S} that it translates to along with the corresponding translation probabilities. For each word $w^T \in T_{w^S}$, we find the concept G_j^T that contains w^T and add a weight p on the edge between the vertex G_i^S and G_j^T, where p is the probability of w^S being translated to w^T. After repeating this process for all the source concepts, we normalize the edge weights such that for each G_i^S, the sum of weights corresponding to the edges connecting G_i^S and any concept in the target language equals to 1. Thus for each source concept the normalized bipartite graph contains a distribution over the target concepts. We call this normalized bipartite graph as the semantic mapping between the two languages. Note, that the normalized bipartite graph can be seen as a matrix M_{map} where the rows and columns correspond to source and target concepts respectively and the entry (i, j) denotes the probability that the i^{th} source concept corresponds to j^{th} target concept.

Now using the matrix M_{map}, the similarity function $Sim^X(d_i^S, d_j^T)$ can be defined as follows:

$$Sim^X(d_i^S, d_j^T) = (\bar{c}_i^S)' M_{map} \bar{c}_j^T \qquad (6)$$

Here, $\bar{c}_i^S$ and $\bar{c}_i^T$ denote the concept vector representation of d_i^S and d_j^T respectively. The concept vector for a document is obtained by replacing the occurrence of each word w_i in the document by its concept.

6 Experimental Evaluation

There are three key questions for which we seek an answer through our experimental evaluation. First, whether the availability of labeled data in a source language is helpful for labeling unlabeled documents in the target language. Second, whether discovery of concepts and concept mapping between languages improves the CLGC performance. Third, given that the target language contains exactly the same classes as the source language (which is not an assumption for CLGC), whether labeling documents using CLGC gives comparable performance to computationally more expensive method that uses a machine translation system. We next describe the dataset, baselines and evaluation metrics that we use to answer these questions.

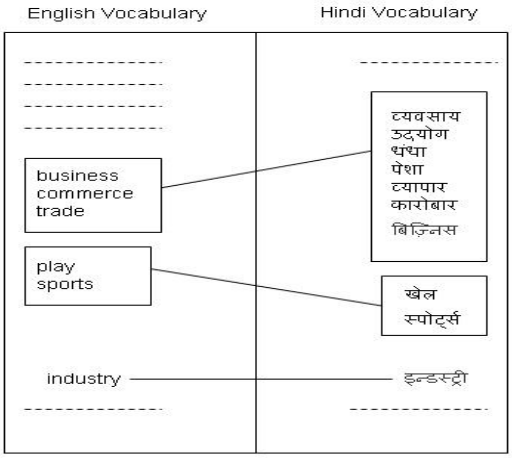

Figure 3: Vocabulary after Semantic Projection

Dataset and Resources: To evaluate the performance of our method, we constructed a dataset of news articles by crawling an English and a Hindi news site. The crawled news articles are from a four month period and belong to the following five categories, *viz*, (1) Economy and Finance - these are news reports on macro-economic events (such as cuts in interest rates, stock market and increase in taxes), (2) Healthcare and BioTech - these are business reports from the Healthcare and Biotechnology industry (mergers and acquisition, patents lawsuits , expansion etc), (3) Energy - these are news reports from the energy and utility sector, (4) sports and (5) Auto. The number of documents for each language and category are shown in Table 1. As mentioned earlier, the CLGC method does not assume that the same set of categories are present in both the languages, to verify this claim we have an additional category, viz, "Auto" in our Hindi dataset which is absent in the English dataset. Even though both English and Hindi news articles are from the same time frame these articles are not aligned.

Language	Economy	BioTech	Energy	Sports	Auto
English	1012	510	500	268	0
Hindi	412	300	350	275	153

Table 1: News Dataset used for Experimentation

	English	Hindi
Number of Unique Words	18128	14521
General Dictionary coverage	11061 (61%)	9344 (64%)
Domain Dictionary coverage	14969 (82.5%)	11767 (81%)

Table 2: Dictionary Statistics

Dictionary	Method	F1	Purity
	K-Means	0.45	0.61
General dictionary	PB	0.49	0.63
	WPB	0.56	0.66
	SM	0.62	0.71
Domain Dictionary	PB	0.57	0.64
	WPB	0.61	0.69
	SM	0.64	0.73

Table 3: Comparison of k means with CLGC using different cross lingual similarity measures

In our experiments, we use an English-Hindi statistical dictionary which was built using the Moses toolkit (Koehn 2007). The training data for the dictionary was a collection of 150,000 English and Hindi parallel sentences sourced from a general corpus. The dictionary built using this corpus is referred to as a "general dictionary" (GD). We further collected 10,000 parallel sentences on the topics present in our news dataset. These were then used along with the earlier set of parallel sentence to learn a dictionary that contains domain specific words and their translations. We refer to this dictionary as a "domain dictionary" (DD). The statistics for these dictionaries in terms of word coverage is shown in Table 2. The objective of creating these two dictionaries is to observe the performance of CLGC when a general purpose dictionary is used in contrast to a domain specific dictionary.

Baselines: One of the objective of experimental evaluation is to see if the availability of source classes helps in clustering documents in the target language. In order to measure gains achieved by the availability of source class information, we compare the performance of CLGC against the standard k-means algorithm. We refer to this as *k-means baseline*.

Another objective of the experimental evaluation is to see whether labeling documents using CLGC gives comparable performance to computationally more expensive method that uses a machine translation system. For this we train a classifier using the English news articles referred to as source classifier. We then translate Hindi new articles into English using Google's machine translation system and then label them using the source classifier. We refer to this as *NB baseline*.

Evaluation Metric The objective of the CLGC approach is to label the unlabeled target dataset. We use the following approach for evaluating this. As the true class-labels for the target news articles are known we assign to each cluster the class-label which is the most frequent in the cluster. All articles in the cluster are now labeled with the corresponding cluster-label. Based on this labeling strategy and the available ground truth we report the accuracy/purity measure which is computed by dividing the correctly labelled documents by the total number of documents. We also evaluate clustering quality by considering the correctness of clustering decisions over all document pairs. We report the standard F1 measure over the pairwise clustering decisions. The F1 measure is the harmonic mean of precision and recall over pairwise decisions.

Experiment 1: In our first experiment, We compare the performance of k-means with the projection based method, referred to as PB, weighted projection based method referred to as WPB and semantic mapping based method, referred to as SM. For this experiment we use the English dataset as the source dataset and Hindi dataset as the target dataset with 4 and 5 categories respectively. For the semantic mapping based method, we discover concepts using the word clustering. The word clustering algorithm uses k-means algorithm. We set k to a large value (we set it to 1000) and use only the first 100 best clusters where goodness of a cluster is measured in terms of its divergence. For each word that is not covered by the first best 100 clusters, we create singleton clusters for the word. We use this procedure for both the source and target dataset. We then use the method described in Section 5 to discover concept mappings.

Since the results obtained for both the k means and all the variations of CLGC depends on the choice of initial centroids, in each experimental run all the methods are seeded with the same set of centroids. The reported results are averaged

over 10 runs with random initialization. We set the value of k equal to the actual number of categories in each dataset for both k-means as well as for CLGC. The value of α in Equation 4 is set to 0.5 and value of n and m in the procedure given in the Figure 2 is kept 20 and 5 respectively.

The results are reported in Table 3. The results show that there is a significant gain that is achieved by CLGC methods over K-means. This shows that the presence of labeled data in the source language helps in the clustering of documents in the target language. We further note that the SM methods, both using "general dictionary" (GD) and "domain dictionary" (DD) outperforms all other methods in their class. This happens because words that do not get translated using the statistical dictionary, are taken into account as they become part of concept mappings that have correspondence across languages. Thus, these terms get accounted in the computation of the SM similarity measure. These terms were not being considered in the PB and WPB similarity computations. As an example the statistical dictionary did not have the translation for the word "bharti" , which is the name of a company from the telecommunication and retail sector. However the word "bharti" mapped to a concept from the source language which contained words such as "communication", "retail" and "ipo". This cluster mapped to a concept in Hindi which had words such as "सँचार", "रिटेल" and "भारती" where the first two words are translations for the words "communication" and "retail" respectively. As a result of this correspondence between the two concepts the words "bharti" and "भारती" get associated. Another key point to note is that the performance of Semantic Mapping using General Dictionary is only slightly worse than Semantic Mapping using the Domain Dictionary. This shows that the semantic mapping based method is able to achieve good performance even when it does not have access to a domain specific dictionary.

Experiment 2: In our second experiment, we compare the performance of SM method which is the best performing CLGC method with the NB baseline. We use the rainbow package (McCallum 1996) to train a naïve Bayes classifier using the English dataset. For translating Hindi documents to English, we use Google [1] translation engine. The accuracy results for this experiment are provided in Table 4.

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

Method	Accuracy
NB	0.71
SM	0.73

Table 4: Comparison of naïve Bayes with CLGC (SM using General Dictionary)

We note that the performance of SM is slightly higher than the naïve Bayes approach. We investigated the reasons behind this and found that there are a few important features that are specific to the Hindi dataset. As the naïve Bayes classifier is trained using the English dataset only, it does not have access to these features and therefore incorrectly classifies the documents that contain such features. While classification techniques such as those based on Support Vector Machines can be expected to perform better than simple NB, our aim here is only to demonstrate that in a resource poor language, where building such classifiers may not be possible (due to the lack of a good machine translation system etc), CLGC can prove to be a useful method.

7 Concluding Remarks

In this paper, we presented cross language guided clustering (CLGC) that utilizes the labeled data from a source language to label unlabeled data from a target language. CLGC tries to cluster unlabeled target language documents such that the resulting clusters are most 'similar' or best 'aligned' to the given source language classes. To achieve this alignment we defined a cross-language similarity measures that returns a similarity score between two documents in different languages. We presented and compared three cross-language similarity measure namely Projection Based, Weighted Projection Based and Semantic Mapping and demonstrate their effectiveness on real-world data-sets. Our Semantic Mapping method, which discovers concepts and their associated mapping across languages, shows the maximum gain in the accuracy of labeling documents over the baseline methods.

References

Xiaojun Wan. 2009. *Co-Training for Cross-Lingual Sentiment Classification*, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics, pages 235–243.

Peter Prettenhofer and Benno Stein. 2010. *Cross-Language Text Classification using Structural Correspondence Learning*. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1118–1127.

Verginica Barbu Mititelu and Radu Ion. 2005. *Automatic Import of Verbal Syntactic Relations Using Parallel Corpora*. Cross-Language Knowledge Induction Workshop.

Indrajit Bhattacharya and Shantanu Godbole and Sachindra Joshi and Ashish Verma. 2009. *Cross-Guided Clustering: Transfer of Relevant Supervision across Domains for Improved Clustering*. Proceedings of the Internationl Conference on Data Mining, pages 41–50.

Mark Hall and Eibe Frank and Geoffrey Holmes and Bernhard Pfahringer and Peter Reutemann and Ian H. Witten. 2009. *The WEKA Data Mining Software: An Update*. SIGKDD Explorations, Volume 11, Issue 1.

Kathleen Mckeown and Regina Barzilay and John Chen and David Elson and David Evans and Judith Klavans and Ani Nenkova and Barry Schiffman and Sergey Sigelman. 2003. *Columbias newsblaster: New features and future directions*. In Proceedings of NAACL-HLT03.

M. Bilenko and R. J. Mooney. 2003 *Adaptive duplicate detection using learnable string similarity measures* In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.

D. Klein and S. D. Kamvar and C. Manning. 2002 *From instance level constraints to space-level constraints: Making the most of prior knowledge in data clustering* In International Conference on Machine Learning, 2002.

I. Dhillon and S. Mallela and D. S. Modha. 2003 *Information theoretic co-clustering* On ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.

N. Slonim and N. Tishby. 2000 *Document clustering using word clusters via the information bottleneck method* In The Annual International ACM SIGIR Conference, 2000.

I. Bhattacharya and L. Getoor. 2007 *Collective entity resolution in relational data* ACM Transactions on Knowledge Discovery from Data, vol. 1, no. 1, pp. 1–36, March 2007.

H. W. Kuhn. 1955. *The hungarian method for the assignment problem* Naval Research Logistics Quarterly, vol. 2, pp. 83–97, 1955.

J. Scott Olsson and Douglas W. Oard and Jan Hajic. 2005. *Cross language text classification* In Proceedings of SIGIR-05, pages 645–646.

Andrew Kachites McCallum 1996. *Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering.* http://www.cs.cmu.edu/ mccallum/bow.

Susan T. Dumais, Todd A. Letsche, Michael L. Littman, and Thomas K. Landauer. 1997. *Automatic cross-language retrieval using latent semantic indexing* In AAAI Symposium on Cross-Language Text and Speech Retrieval.

Blaz Fortuna and John Shawe-Taylor. 2005. *The use of machine translation tools for cross-lingual text mining.* In Proceedings of the ICML Workshop on Learning with Multiple Views.

Ke Wu and Bao-Liang Lu. 2007. *Cross-Lingual Document Clustering* . In Lecture Notes in Computer Science, Volume 4426/2007, 956–963,

Philipp Koehn, Hieu Hoang, Alexandra Birch Mayne, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst 2007. *Open source toolkit for statistical machine translation.* Annual Meeting of the Association for Computation Linguistics (ACL), Demonstration Session