Cross-Lingual Sentiment Analysis using modified BRAE

Sarthak Jain
Department of Computer Engineering
Delhi Technological University
DL, India
successar@gmail.com

Shashank Batra
Department of Computer Engineering
Indian Institute of technology, Delhi
DL, India
shashankg@gmail.com

Abstract

Cross-Lingual Learning provides a mechanism to adapt NLP tools available for label rich languages to achieve similar tasks for label-scarce languages. An efficient cross-lingual tool significantly reduces the cost and effort required to manually annotate data. In this paper, we use the Recursive Autoencoder architecture to develop a Cross Lingual Sentiment Analysis (CLSA) tool using sentence aligned corpora between a pair of resource rich (English) and resource poor (Hindi) language. The system is based on the assumption that semantic similarity between different phrases also implies sentiment similarity in majority of sentences. The resulting system is then analyzed on a newly developed Movie Reviews Dataset in Hindi with labels given on a rating scale and compare performance of our system against existing systems. It is shown that our approach significantly outperforms state of the art systems for Sentiment Analysis, especially when labeled data is scarce.

1 Introduction

Sentiment Analysis is a NLP task that deals with extraction of opinion from a piece of text on a topic. This is used by a large number of advertising and media companies to get a sense of public opinion from their reviews. The ever increasing user generated content has always been motivation for sentiment analysis research, but majority of work has been done for English Language. However, in recent years, there has been emergence of increasing amount of text in Hindi on electronic sources but NLP Frameworks to process this data is sadly miniscule. A major cause for this is the lack of annotated datasets in Indian Languages.

One solution is to create cross lingual tools between a resource rich and resource poor language that exploit large amounts of unlabeled data and sentence aligned corpora that are widely available on web through bilingual newspapers, magazines, etc. Many different approaches have been identified to perform Cross Lingual Tasks but they depend on the presence of MT-System or Bilingual Dictionaries between the source and target language.

In this paper, we use Bilingually Constrained Recursive Auto-encoder (BRAE) given by (Zhang et al., 2014) to perform Cross Lingual sentiment analysis. Major Contributions of this paper are as follows: First, We develop a new Rating scale based Movie Review Dataset for Hindi. Second, a general framework to perform Cross Lingual Classification tasks is developed by modifying the architecture and training procedure for BRAE model. This model exploits the fact that phrases in two languages, that share same semantic meaning, can be used to learn language independent semantic vector representations. These embeddings can further be fine-tuned using labeled dataset in English to capture enough class information regarding Resource poor language. We train the resultant framework on English-Hindi Language pair and evaluate it against state of the art SA systems on existing and newly developed dataset.

2 Related Work

2.1 Sentiment Analysis in Hindi

In recent years, there have been emergence of works on Sentiment Analysis (both monolingual and cross-lingual) for Hindi. (Joshi et al., 2010) provided a comparative analysis of Unigram based In-language, MT based Cross Lingual and WordNet based Sentiment classifier, achieving highest accuracy of 78.14%. (Mittal et al., 2013) described a system based on Hindi SentiWordNet for assign-

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 159–168,
Lisbon, Portugal, 17-21 September 2015.

ing positive/negative polarity to movie reviews. In this approach, overall semantic orientation of the review document was determined by aggregating the polarity values of the words in the document assigned using the WordNet. They also included explicit rules for handling Negation and Discourse relations during preprocessing in their model to achieve better accuracies.

For Languages where labeled data is not present, approaches based on cross-lingual sentiment analysis are used. Usually, such methods need intermediary machine translation system (Wan et al., 2011; Brooke et al., 2009) or a bilingual dictionary (Ghorbel and Jacot, 2011; Lu et al., 2011) to bridge the language gap. Given the subtle and different ways in which sentiments can be expressed and the cultural diversity amongst different languages, an MT system has to be of a superior quality to perform well(Balamurali et al., 2012).

(Balamurali et al., 2012) present an alternative approach to Cross Lingual Sentiment Analysis (CLSA) using WordNet senses as features for supervised sentiment classification. A document in Resource Poor Language was tested for polarity through a classifier trained on sense marked and polarity labeled corpora in Resource rich language. The crux of the idea was to use the linked WordNets of two languages to bridge the language gap.

Recently, (Popat et al., 2013) describes a Cross Lingual Clustering based SA System. In this approach, features were generated using syntagmatic property based word clusters created from unlabeled monolingual corpora, thereby eliminating the need for Bilingual Dictionaries. These features were then used to train a linear SVM to predict positive or negative polarity on a tourism review dataset.

2.2 Autoencoders in NLP Tasks

Autoencoders are neural networks that learn a low dimensional vector representation of fixed-size inputs such as image segments or bag-of-word representations of documents. They can be used to efficiently learn feature encodings that are useful for classification. The Autoencoders were first applied in a recursive setting by Pollack (1990) in recursive auto-associative memories (RAAMs). However, RAAMs needed fixed recursive data structures to learn vector representations, whereas RAE given by (Socher et al., 2011) builds recursive data structure using a greedy algorithm. The RAE can be pre-trained with an unsupervised algorithm and then fine-tuned according to the label of the phrase, such as the syntactic category in parsing(Socher et al., 2013), the polarity in sentiment analysis, etc. The learned structures are not necessarily syntactically accurate but can capture more of the semantic information in the word vectors.

3 BRAE Framework

(Zhang et al., 2014) used the RAE along with a Bilingually Constrained Model to simultaneously learn phrase embeddings for two languages in semantic vector space. The core idea behind BRAE is that a phrase and its correct translation should share the same semantic meaning. Thus, they can supervise each other to learn their semantic phrase embeddings. Similarly, non-translation pairs should have different semantic meanings, and this information can also be used to guide learning semantic phrase embeddings. In this method, a standard recursive autoencoder (RAE) pre-trains the phrase embedding with an unsupervised algorithm by greedily minimizing the reconstruction error (Socher et al., 2011), while the bilingually-constrained model learns to finetune the phrase embedding by minimizing the semantic distance between translation equivalents and maximizing the semantic distance between non-translation pairs.

In this section, We will briefly present the structure and training algorithm for BRAE model. After that, we show how this model can be adapted to perform CLSA.

3.1 Recursive Auto-encoder Framework

In this model, each word w_k in the vocabulary V of given language corresponds to a vector $x_k \in \mathbb{R}^n$ and stacked into a single word embedding matrix $L \in \mathbb{R}^{n \times |V|}$. This matrix is learned using DNN (Collobert and Weston, 2008; Mikolov et al., 2013) and serves as input to further stages of RAE.

Using this matrix, a phrase $(w_1 w_2 \dots w_m)$ is first projected into a list of vectors $(x_1, x_2, \dots x_m)$. The RAE learns the vector representation of the phrase by combining two children vectors recursively in a bottom-up manner. For two children $c_1 = x_1, c_2 = x_2$, the auto-encoder computes the parent vector y_1:

$$y_1 = f(W^{(1)}[c_1; c_2] + b^{(1)}); y_1 \in \mathbb{R}^n \quad (1)$$

To assess how well the parent vector represents its children, the auto-encoder reconstructs the chil-

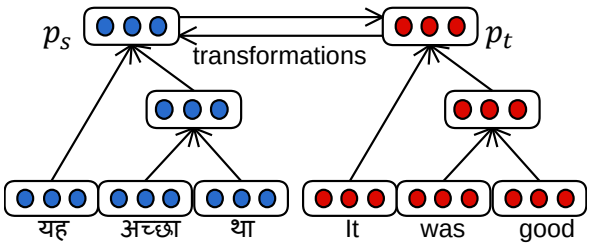

Figure 1: An illustration of BRAE structure

dren :

$$[c_1'; c_2'] = W^{(2)}p + b^{(2)} \quad (2)$$

and tries to minimize the reconstruction error (Euclidean Distance) $E_{rec}([c_1; c_2])$ between the inputs $[c_1; c_2]$ and their reconstructions $[c_1'; c_2']$.

Given y_1, Eq.1 is used again to compute y_2 by setting the children to be $[c_1; c_2] = [y_1; x_3]$. The same auto-encoder is re-used until the vector of the whole phrase is generated. For unsupervised phrase embedding, the sum of reconstruction errors at each node in binary tree y is minimized:

$$E_{rec}(x; \theta) = \arg\min_{y \in A(x)} \sum_{k \in y} E_{rec}([c_1; c_2]_k) \quad (3)$$

Where $A(x)$ denotes all the possible binary trees that can be built from inputs x. A greedy algorithm is used to generate the optimal binary tree y^*. The parameters $\theta^{rec} = (\theta^{(1)}, \theta^{(2)})$ are optimized over all the phrases in the training data. For further details, please refer (Socher et al., 2011)

3.2 Semantic Error

The BRAE model jointly learns two RAEs for source language L_S and target language L_T. Each RAE learn semantic vector representation p_s and p_t of phrases s and t respectively in translation-equivalent phrase pair (s, t) in bilingual corpora (shown in Fig.1). The transformation between the two is defined by:

$$p_t' = f(W_s^t p_s + b_s^t), p_s' = f(W_t^s p_t + b_t^s) \quad (4)$$

where $\theta_s^t = (W_s^t, b_s^t), \theta_t^s = (W_t^s, b_t^s)$ are new parameters introduced.

The semantic error between learned vector representations p_s and p_t is calculated as :

$$E_{sem}(s, t; \theta) = E_{sem}^*(t|s; \theta_t^s) + E_{sem}^*(s|t; \theta_s^t) \quad (5)$$

where $E_{sem}^*(s|t; \theta_t^s)$ is the semantic distance of p_s given p_t and vice versa. To calculate it, we first calculate Euclidean distance between original p_t and transformation p_t' as $D_{sem}(s|t, \theta_s^t) = \frac{1}{2}\|p_t - p_t'\|^2$. The max-semantic-margin distance between them is then defined as

$$E_{sem}^*(s|t, \theta_s^t) = max\{0, D_{sem}(s|t, \theta_s^t) - D_{sem}(s|t', \theta_s^t) + 1\} \quad (6)$$

where we simultaneously minimize the distance between translation pairs and maximized between non-translation pairs. Here t' in non-translation pair (s, t') is obtained by replacing the words in t with randomly chosen target language words. We calculate the $E_{sem}^*(t|s; \theta_t^s)$ in similar manner.

3.3 BRAE Objective Function

Thus, for the phrase pair (s, t), the joint error becomes:

$$\begin{aligned} E(s, t, \theta) &= E(s|t, \theta) + E(t|s, \theta) \\ E(s|t, \theta) &= \alpha E_{rec}(s; \theta_s^{rec}) + (1 - \alpha) E_{sem}^*(s|t, \theta_s^t)) \\ E(t|s, \theta) &= \alpha E_{rec}(t; \theta_t^{rec}) + (1 - \alpha) E_{sem}^*(t|s, \theta_t^s)) \end{aligned} \quad (7)$$

The hyper-parameter α weighs the reconstruction and semantic errors. The above equation indicates that the Parameter sets $\theta_t = (\theta_t^s, \theta_t^{rec})$ and $\theta_s = (\theta_s^t, \theta_s^{rec})$ on each side respectively can be optimized independently as long as the phrase representation of other side is given to compute semantic error.

The final BRAE objective over the phrase pairs training set (S, T) becomes:

$$J_{BRAE} = \frac{1}{N} \sum_{(s,t) \in (S,T)} E(s, t; \theta) + \frac{\lambda_{BRAE}}{2} \|\theta\|^2 \quad (8)$$

3.4 Unsupervised Training of BRAE

The word embedding matrices L_s and L_t are pre-trained using unlabeled monolingual data with Word2Vec toolkit (Mikolov et al., 2013). All other parameters are initialized randomly. We use SGD algorithm for parameter optimization. For full gradient calculations for each parameter set, please see (Zhang et al., 2014).

1. RAE Training Phase: Apply RAE Framework (Sec. 3.1) to pre-train the source and target phrase representations p_s and p_t respectively by optimizing θ_s^{rec} and θ_t^{rec} using unlabeled monolingual datasets.

2. Cross-Training Phase: Use target-side phrase representation p_t to update the source-side

parameters θ_s and obtain source-side phrase representation p'_s, and vice-versa for p_s. Calculate the joint error over the bilingual training corpus. On reaching a local minima or predefined no. of iterations (30 in our case), terminate this phase, otherwise set $p_s = p'_s$, $p_t = p'_t$, and repeat.

4 Adapting Model for Classifying Sentiments

At the end of previous Training procedure, we obtain high quality phrase embeddings in both source and target language and transformation function between them. We now extend that model to perform cross lingual supervised tasks, specifically CLSA.

To achieve this, we need to modify the learned semantic phrase embeddings such that they can capture information about sentiment. Since we only use monolingual labeled datasets from this point onwards, the supervised learning phases will occur independently for each RAE as we do not have any "phrase pairs" now. Thus, the new semantic vector space generated for word and phrase embeddings may no longer be in sync with their corresponding transformations.

We propose following modifications to the system to deal with this problem. Let L_S and L_T represent Resource rich and Resource poor language respectively in above model.

Modifications in architecture: We first include a softmax (σ) layer on top of each parent node in RAE for L_S to predict a K-dimensional multinomial distribution over the set of output classes defined by the task (e.g : polarity, Ratings).

$$d(p; \theta^{ce}) = \sigma(W^{ce} p) \tag{9}$$

Given this layer, we calculate cross entropy error $E_{ce}(p_k, t, W_{ce})$ generated for node p_k in binary tree, where t is target multinomial distribution or one-hot binary vector for target label. We use this layer to capture and predict actual sentiment information about the data in both L_S and L_T (described in next section). We show a node in modified architecture in Fig.2.

Penalty for Movement in Semantic Vector space: During subsequent training phases, we include the euclidean norm of the difference between the original and new phrase embeddings as penalty in reconstruction error at each node of the tree.

$$E^*_{rec}([c_1; c_2]; \theta) = E_{rec}([c_1; c_2]; \theta) + \frac{\lambda_p}{2} \|p - p^*\|^2 \tag{10}$$

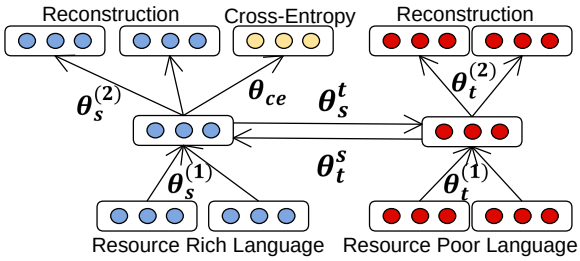

Figure 2: An illustration of BRAE segment with Cross Entropy layer

Here p is the phrase representation we get during forward propagation of current training iteration and p^* is the representation we get if we apply the parameters obtained at the end of the Cross training phase to children $[c_1; c_2]$ of that node. The reason to do this is twofold.

First, during supervised training, the error will back propagate through RAEs for both languages affecting their respective weights matrices and word embeddings. This will modify the semantic representation of phrases captured during previous phases of training procedure and adversely affect the transformations derived from them. Therefore we need to include some procedure such that the transformation information learned during Cross-training phase is not lost.

Secondly, we observe that the information about the semantic similarity of a word or phrase also implies sentiment similarity between the two. That is when dealing with bilingual data, words or phrases that appear near each other in semantic space typically represent common sentiment information and we want our model to create a decision boundary around these vectors instead of modifying them too much.

Disconnecting the RAEs: We fix the transformation weights between the two RAEs, i.e. in subsequent training steps the transformation weights(θ^t_s, θ^s_t) are not modified but rather pass the back propagated error as it is to previous layers. We observed that on optimizing the objective along with the penalty term, the transformation weights are preserved between new semantic/sentiment vector spaces, resulting in slightly degraded performance, but were still able to preserve enough information about the semantic structure of two languages.Also, it reinforced the penalty imposed on the movement of phrase embeddings in semantic vector space.On the other hand, if the weights were allowed to be updated,

the accuracies were affected severely as information learned during previous phases was lost and the weights were not been able to capture enough information about the modified phrase embeddings and generalize well on test phrases not encountered in labeled training set of Resource Scarce Language.

4.1 Supervised Training Phases

We now explain supervised training procedure using only monolingual labeled data for each language. These training phases occur at the end of BRAE training. In each training phase, we use SGD algorithm to perform parameter optimization.

4.1.1 Phase I : Resource Rich language

In this phase, we only modify the parameters of RAE_{L_S}, i.e. θ_s^{rec} and θ_{ce} by optimizing following objective over (sentence, label) pairs (x, t) in its labeled corpus.

$$J_S = \frac{1}{N} \sum_{(x,t)} E(x, t; \theta) + \frac{\lambda_S}{2} \|\theta\|^2 \qquad (11)$$

where $E(x, t; \theta)$ is the sum over the errors obtained at each node of the tree that is constructed by the greedy RAE:

$$E(x, t; \theta) = \sum_{k \in RAE_{L_S}(x)} \kappa E^*_{rec}([c_1; c_2]_k; \theta_s) + (1 - \kappa) E_{ce}(p_k, t; \theta_{ce}) \qquad (12)$$

To compute this gradient, we first greedily construct all trees and then derivatives for these trees are computed efficiently via back-propagation through structure (Goller and Kuchler, 1996). The gradient for our new reconstruction function (Eq. 10) w.r.t to p at a given node is calculated as

$$\frac{\partial E^*_{rec}}{\partial p} = \frac{\partial E_{rec}}{\partial p} + \lambda_p (p - p^*) \qquad (13)$$

The first term $\frac{\partial E_{rec}}{\partial p}$ is calculated as in standard RAE model. The partial derivative in above equation is used to compute parameter gradients in standard back-propagation algorithm.

4.1.2 Phase II : Resource Poor Language

In this phase, we modify the parameters of RAE_{L_T} and θ_{ce} by optimizing Objective J_T over (sentence, label) pairs (x, t) in labeled corpus for L_T (much smaller than that for L_S). The equation for J_T is similar to Eq.11 and Eq.12 but with θ_t and η as parameters instead of θ_s and κ respectively.

Since cross-entropy layer is only associated with L_S, we need to traverse the transformation parameters to obtain sentiment distribution for each node (green path in Fig.2). That is, we first transform p_t to source side phrase p'_s and then apply the cross entropy weights to it.

$$d(p_t, \theta_{ce}) = \sigma(W^{ce}.f(W^t_s p_t + b^t_s)) \qquad (14)$$

We use the similar back-propagation through structure approach for gradient calculation in Phase I. During back propagation, 1) we do not update the transformation weights, 2) we transfer error signals during back-propagation from Cross-entropy layer to $\theta^{(1)}_t$ as if the transformation was an additional layer in the network.

4.1.3 Predicting overall sentiment

To predict overall sentiment associated with the sentence in L_T, we use the phrase embeddings p_t of the top layer of the RAE_{L_T} and it transformation p'_s. Together, we train a softmax regression classifier on concatenation of these two vector using weight matrix $W \in \mathbb{R}^{K \times 2n}$

5 Experimental Work

We perform experiments on two kind of sentiment analysis systems : (1) that gives +ve/-ve polarity to each review and (2) assigns ratings in range 1 - 4 to each review.

5.1 External Datasets Used

For pre-training the word embeddings and RAE Training, we used HindMonoCorp 0.5(Bojar et al., 2014) with 44.49M sentences (787M Tokens) and English Gigaword Corpus.

For Cross Training, we used the bilingual sentence-aligned data from HindEnCorp[1] (Bojar et al., 2014) with 273.9k sentence pairs (3.76M English, 3.88M Hindi Tokens). This dataset contains sentence pair obtained from Bilingual New Articles, Wikipedia entries, Automated Translations, etc. Training and Validation division is 70% and 30% for all above datasets.

In Supervised Phase I, we used IMDB11 dataset available at *http://ai.stanford.edu/~amaas/data/sentiment/* and first used by (Maas et al., 2011) for +ve/-ve

[1] http://ufal.mff.cuni.cz/hindencorp

system containing 25000 +ve and 25000 -ve movie reviews.

For 4-ratings system, we use Rotten Tomatoes Review dataset (scale dataset v1.0) found at *http://www.cs.cornell.edu/People/pabo/movie-review-data*. The dataset is divided into four author-specific corpora, containing 1770, 902, 1307, and 1027 documents and each document has accompanying 4-Ratings ({0, 1, 2, 3}) label.

5.2 Rating Based Hindi Movie Review (RHMR) Dataset

We crawled the Hindi Movie Reviews Website[2] to obtain 2945 movie reviews. Each Movie Review on this site is assigned rating in range 1 to 4 by at least three reviewers. We first discard reviews that whose sum of pairwise difference of ratings is greater than two. The final rating for each review is calculated by taking the average of the ratings and rounding up to nearest integer. The fraction of Reviews obtained in ratings 1-4 are [0.20, 0.25, 0.35, 0.20] respectively. Average length of reviews is 84 words. For +ve/-ve polarity based system, we group the reviews with ratings {1, 2} as negative and {3, 4} as positive.

5.3 Experimental Settings

We used following Baselines for Sentiment Analysis in Hindi :

Majority class: Assign the most frequent class in the training set (Rating:3 / Polarity:+ve)

Bag-of-words: Softmax regression on Binary Bag-of-words

We also compare our system with state of the art Monolingual and Cross Lingual System for Sentiment Analysis in Hindi as described by (Popat et al., 2013) using the same experimental setup. The best systems in each category given by them are as below:

WordNet Based: Using Hindi-SentiWordNet[3], each word in a review was mapped to corresponding synset identifiers. These identifiers were used as features for creating sentiment classifiers based on Binary/Multiclass SVM trained on bag of words representation using libSVM library.

Cross Lingual (XL) Clustering Based: Here, joint clustering was performed on unlabeled bilingual corpora which maximizes the joint likelihood of monolingual and cross-lingual factors.. For details, please refer the work of (Popat et al., 2013). Each word in a review was then mapped to its cluster identifier and used as features in an SVM.

Our approaches

Basic RAE: We use the Semi-Supervised RAE based classification where we first trained a standard RAE using Hindi monolingual corpora, then applied supervised training procedure as described in (Socher et al., 2011). This approach doesn't use bilingual corpora, but is dependent on amount of labeled data in Hindi.

BRAE-U: We neither include penalty term, nor fix the transformations weights in our proposed system.

BRAE-P: We only include the penalty term but allow the transformation weights to be modified in proposed system.

BRAE-F: We add the penalty term and fix the transformation weights during back propagation in proposed system.

5.4 Experimental Setup

We combined the text data from all English Datasets (English Gigaword + HindEnCorp English Portion + IBMD11 + Scale Dataset) described above to train the word embeddings using Word2Vec toolkit and RAE. Similarly, we combined text data from all Hindi Datasets (HindMonoCorp + HindiEnCorp Hindi Portion + RHMR) to train word embeddings and RAE for Hindi.

We used MOSES Toolkit (Koehn et al., 2007) to obtain high quality bilingual phrase pairs from HindEnCorp to train our BRAE model. After removing the duplicates, 364.3k bilingual phrase pairs were obtained with lengths ranging from 1-6, since bigger phrases reduced the performance of the system in terms of Joint Error of BRAE model.

We randomly split our RHMR dataset into 10 segments and report the average of 10-fold cross validation accuracies for each setting for both Ratings and Polarity classifiers.

We also report 5-fold cross validation accuracy on Standard Movie Reviews Dataset (hereby referred as SMRD) given by (Joshi et al., 2010) which contains 125 +ve and 125 -ve reviews in Hindi. The dataset can be obtained at http://www.cfilt.iitb.ac.in/Resources.html.

Since this project is about reducing dependence on annotated datasets, we experiment on how accuracy varies with labeled training dataset (RHMR) size. To perform this, we train our model

[2]http://hindi.webdunia.com/bollywood-movie-review/

[3]http://www.cfilt.iitb.ac.in/

in 10% increments (150 examples) of training set size (each class sampled in proportion of original set). For each size, we sample the data 10 times with replacement and trained the model. For each sample, we calculated 10-fold cross validation accuracy as described above. Final accuracy for each size was calculated by averaging the accuracies obtained on all 10 samples. Similar kind of evaluation is done for all other Baselines explored.

In subsequent section, the word 'significant' implies that the results were statistically significant ($p < 0.05$) with paired T-test

5.5 BRAE Hyper Parameters

We empirically set the learning rate as 0.05. The word vector dimension was selected as 80 from set $[40, 60, 80, 100, 120]$ using Cross Validation. We used joint error of BRAE model to select α as 0.2 from range $[0.05, 0.5]$ in steps of 0.05. Also, λ_L was set as 0.001 for DNN trained for word embedding and λ_{BRAE} as 0.0001.

For semi-supervised phases , we used 5-fold cross validation on training set to select κ and η in range $[0.0, 1.0]$ in steps of 0.05 with optimal value obtained at $\kappa = 0.2$ and $\eta = 0.35$. Parameter λ_p was selected as 0.01 , λ_S as 0.1 and λ_T as 0.04 after selection in range $[0.0, 1.0]$ in steps of 0.01.

5.6 Results

Dataset	RHMR		SMRD
Classifier	**Ratings**	**Polarity**	**Polarity**
Majority class	35.19	51.83	52.34
Bag-of-Words	51.98	62.52	68.47
WordNet based	55.47	67.29	75.5
XL Clustering	72.34	84.46	84.71
Basic RAE	75.53	79.31	81.06
BRAE-U	76.01	82.66	84.83
BRAE-P	79.70	84.85	87.00
BRAE-F	81.22	90.50	90.21

Table 1: Accuracies obtained for various Experimental Settings. Model are trained on complete labeled training datasets

Table 1 present the results obtained for both ratings based and polarity classifier on RHMR and MRD Dataset. Our model gives significantly better performance for ratings based classification than any other baseline system currently used for SA in Hindi. The margin of accuracy obtained against next best classifier is about 8%. Also, for +ve/-ve polarity classifier, the accuracy showed an improvement of 6% over next highest baseline.

$A \downarrow / P \rightarrow$	P-1	P-2	P-3	P-4
A-1	83.19	15.28	1.53	0.00
A-2	12.23	82.20	5.57	0.00
A-3	0.00	9.03	81.26	9.71
A-4	0.00	1.87	19.69	78.44
F1-score	0.83	0.78	0.82	0.80

Table 2: Confusion Matrix for Ratings by BRAE-F, **Across**: Predicted Rating, **Downward:** Actual Rating

In Table 2, we calculate the confusion matrix for our model(BRAE-F) for the 4-Ratings case. Value in a cell (A_i, P_j) represents the percentage of examples in actual rating class i that are predicted as rating j. We also show the F1 score calculated for each individual rating class. It clearly shows that our model has low variation in F1-scores and thereby its performance among various rating classes.

In Fig. 3, we show the variation in accuracy of the classifiers with amount of sentiment labeled Training data used. We note that our approach consistently outperforms the explored baselines at all dataset sizes. Also, our model was able to attain accuracy comparable to other baselines at about 50% less labeled data showing its strength in exploiting the unlabeled resources.

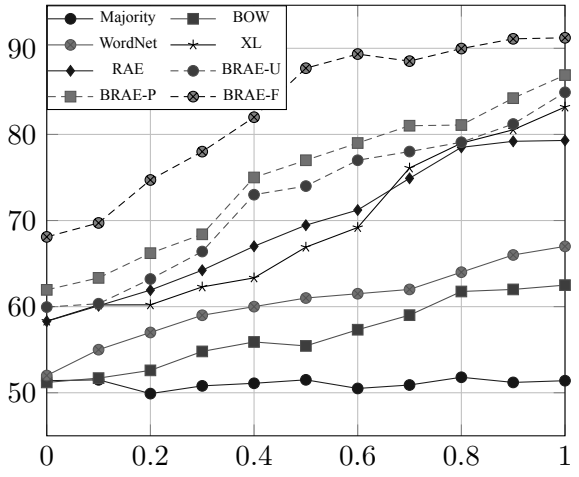

Figure 3: Variation of Accuracy (+ve/-ve Polarity) with Size of labeled Dataset(Hindi), **x-axis:** Fraction of Dataset Used, **y-axis:** %age Accuracy Obtained

We also experiment with variation of accuracies

New Word/Phrase	Similar Words/Phrases	Sentiment label
depressing निराशाजनक	gloomy उदास discouraging निराशात्मक	Rating : 1 Polarity : -ve
was painful दर्दनाक था	was difficult कठिन था was bad खराब था	Rating : 2 Polarity : -ve
should be awarded सम्मानित किया जाना चाहिए	was appreciated सराहना की गई will get accolades वाहवाही मिलना चाहिए	Rating : 4 Polarity : +ve
public won't come लोग नहीं आएगा	no one will come कोई नहीं आएगा viewers won't come दर्शक नहीं आएगा	Rating : 1 Polarity : -ve

Table 3: Semantically similar phrases obtained for new phrases and their assigned label

with amount of Unlabeled Bilingual Training Data used for Cross Lingual models explored. Again we increase size of bilingual dataset in 10% increments and calculate the accuracy as described previously. In Fig. 4, we observed that performance of the proposed approach steadily increases with amount of data added, yet even at about 50000 (20%) phrase pairs, our model produces remarkable gains in accuracy.

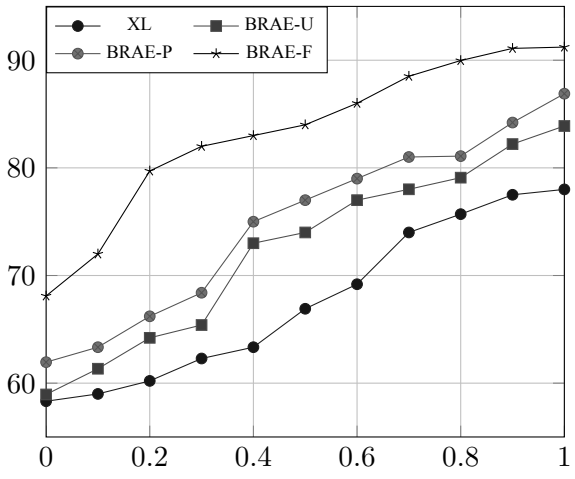

Figure 4: Variation of Accuracy (+ve/-ve polarity) with Size of Unlabeled Bilingual Corpora, **x-axis:** Fraction of Training Data Used, **y-axis:** %age Accuracy Obtained

We also observed that the model which restricts modification to transformation weights during supervised phase II does better than the one which allows the modification at all dataset sizes. This result appears to be counterintuitive to normal operation of neural network based models, but supports our hypothesis as explained in previous sections.

5.7 Performance and Error Analysis

Analysis on the test results showed that the major advantage given by our model occurs due to presence of unknown words (i.e.words not present in labeled dataset) in test data. Since we restricted the movement in semantic vector space, our model was able to infer the sentiment for a unknown word/phrase by comparing it with semantically similar words/phrases. In Table 3, we extracted the Top-2 semantically similar phrases in training set for small new phrases and sentiment labeled assigned to them by our model (the phrases are manually translated from Hindi for reader's understanding). As we can see, our model was able to extract grammatically correct phrases with similar semantic nature as given phrase and assign correct sentiment label to it.

Secondly, We found that our model was able to correctly infer word sense for polysemous words that adversely affected the quality of sentiment classifiers in our baselines. This eliminates the need for manually constructed fine grained lexical resource like WordNets and development of automated annotation resources. For example, to a phrase like *"Her acting of a schizophrenic mother made our hearts **weep"***, the baselines classifiers assigned negative polarity due to presence of words like 'weep', yet our model was correctly able to predict positive polarity and assigned it a rating of 3.

Error Analysis of test results showed that errors made by our model can be classified in two major categories :

1) A review may only give description of the object in question (in our case , the description of the film) without actually presenting any individual sentiments about it or it may express conflicting sentiments about two different aspects about the same object. This presents difficulty in assign-

ing a single polarity/rating to the review.

2) Presence of subtle contextual references affected the quality of predictions made by our classifier. For example, sentence like *"His poor acting generally destroys a movie, but this time it didn't"* got a rating of 2 due to presence of phrase with negative sense (here the phrase doesn't have ambiguous sense), yet the actual sentiment expressed is positive due to temporal dependence and generalization. Also, *"This movie made his last one looked good"* makes a reference to entities external to the review, which again forces our model to make wrong prediction of rating 3.

Analyzing these aspects and making correct predictions on such examples needs further work.

6 Conclusion and Future Work

This study focused on developing a Cross Lingual Supervised Classifier based on Bilingually Constrained Recursive Autoencoder. To achieve this, our model first learns phrase embeddings for two languages using Standard RAE, then fine tune these embeddings using Cross Training procedure. After imposing certain restrictions on these embeddings, we perform supervised training using labeled sentiment corpora in English and a much smaller one in Hindi to get the final classifier.

The experimental work showed that our model was remarkably effective for classification of Movie Reviews in Hindi on a rating scale and predicting polarity using least amount of data to achieve same accuracy as other systems explored. Moreover it reduces the need for MT System or lexical resources like Linked WordNets since the performance is not degraded too much even when we lack large quantity of labeled data.

In Future, we hope to 1) extend this system to learn phrase representations among multiple languages simultaneously, 2) apply this framework to other cross Lingual Tasks such as Paraphrase detection, Question Answering, Aspect Based Opinion Mining etc and 3) Learning different weight matrices at different nodes to capture complex relations between words and phrases.

References

Balamurali, Aditya Joshi, and Pushpak Bhattacharyya. 2012. Cross-lingual sentiment analysis for Indian languages using linked wordnets. In *Proceedings of COLING 2012: Posters*, pages 73--82. The COLING 2012 Organizing Committee.

Ondřej Bojar, Vojtěch Diatka, Pavel Rychlý, Pavel Straňák, Vít Suchomel, Aleš Tamchyna, and Daniel Zeman. 2014. HindEnCorp - Hindi-English and Hindi-only Corpus for Machine Translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA).

Julian Brooke, Milan Tofiloski, and Maite Taboada. 2009. Cross-linguistic sentiment analysis: From english to spanish. In *RANLP*, pages 50--54.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160--167. ACM.

Hatem Ghorbel and David Jacot. 2011. Further experiments in sentiment analysis of french movie reviews. In *Advances in Intelligent Web Mastering--3*, pages 19--28. Springer.

Christoph Goller and Andreas Kuchler. 1996. Learning task-dependent distributed representations by backpropagation through structure. In *Neural Networks, 1996., IEEE International Conference on*, volume 1, pages 347--352. IEEE.

Aditya Joshi, AR Balamurali, and Pushpak Bhattacharyya. 2010. A fall-back strategy for sentiment analysis in hindi: a case study. *Proceedings of the 8th ICON*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177--180. Association for Computational Linguistics.

Bin Lu, Chenhao Tan, Claire Cardie, and Benjamin K Tsou. 2011. Joint bilingual sentiment classification with unlabeled parallel corpora. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 320--330. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142--150. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111--3119.

Namita Mittal, Basant Agarwal, Garvit Chouhan, Nitin Bania, and Prateek Pareek. 2013. Sentiment analysis of hindi review based on negation and discourse relation. In *proceedings of International Joint Conference on Natural Language Processing*, pages 45--50.

Kashyap Popat, Balamurali A.R, Pushpak Bhattacharyya, and Gholamreza Haffari. 2013. The haves and the have-nots: Leveraging unlabelled corpora for sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 412--422. Association for Computational Linguistics.

Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 151--161. Association for Computational Linguistics.

Richard Socher, John Bauer, Christopher D Manning, and Andrew Y Ng. 2013. Parsing with compositional vector grammars. In *In Proceedings of the ACL conference*. Citeseer.

Chang Wan, Rong Pan, and Jiefei Li. 2011. Bi-weighting domain adaptation for cross-language text classification. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1535.

Jiajun Zhang, Shujie Liu, Mu Li, Ming Zhou, and Chengqing Zong. 2014. Bilingually-constrained phrase embeddings for machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 111--121. Association for Computational Linguistics.