International Journal of Advance and Innovative Research
Volume 6, Issue 1 (XXVI): January - March, 2019

ISSN 2394 - 7780

BILINGUALLY CONSTRAINED RECURSIVE AUTO ENCODER (BRAE)

Shilpa G V

ABSTRACT

Sentiment Analysis is a process of extracting one's opinion and analyzing weather the opinion is positive, negative or neutral. Cross Lingual Sentiment Analysis (CLSA) deals with the prediction of the opinion of the content in the testing language using the trained language which is trained with the help of classifiers. CLSA adopts the features of Natural Language Processing (NLP) tools. This provides the mechanism to adopt the similar features present in the resource rich language to the resource poor language. Resource rich language has a good availability of labeled corpus, resource poor language scares with the availability of the labeled corpus. Syntax based features and lexical based features were adopted in the traditional methods of the classification of the Sentiment Analysis. Some of the popular approaches make use of the Machine Translation method. This was helpful in converting the documents of test language to the trained language using the training language classifier. Machine translation method fails for most of the language pairs and when they were successful; its accuracy was too low. This made to follow the other approaches and led to the introduction of Bilingually Constrained Recursive Auto Encoder (BRAE).

INTRODUCTION

Sentiment Analysis has been performed most on the English language, but there is a work in other languages such as German, Chinese, and Spanish etc and also there are other works that has been carried out in the regional language. In order to perform the Sentiment Analysis on these languages Cross Lingual methods has to be used because of the scarcity of the labeled corpus in these languages. Trained language in the CLSA predicts and analyzes the opinion of the testing language. CLSA often made use of Machine Translation method which converted the text in test language to the trained language using the training classifier. But this method could not sustain for the longer time since it failed for the most of the pairs in a language. This led to evolve efficient tool of Cross Lingual Sentiment Analysis which helped reducing the effort of annotating the data manually. This paper introduces the architecture of Recursive Auto-Encoder and Bilingually Constrained Recursive Auto Encoder to perform the CLSA between the resource rich and resource poor language. One language acts as a source language where as other represents the target language. Linked word net of two Indian regional languages has been used to reduce the language gap in between these languages. Hindi and Marathi have been used as two languages which basically do not make use of the Machine Translation method. Based on the survey, there was 72% of accuracy for Hindi and 84% of accuracy for Marathi[1]. When the bilingual dictionary has been used, there is improvement in the accuracy over 13% to 15%. New word net has been created for the target language correspondingly that matches the synset from the source language wordnet with the help of expansion based methods. As a result of this both source and the target language now have same synset identifier.

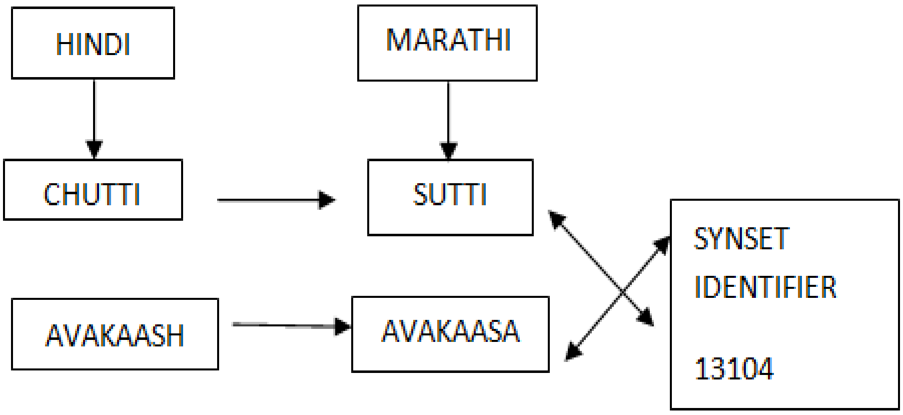

Figure-1: Example of Multidict of Hindi and Marathi

Figure 1 represents that Hindi and Marathi word nets have been build with the expansion based method. In the above figure 1, Hindi and Marathi acts as a resource rich and resource poor language respectively. Here Marathi word net is getting expanded by receiving the word net from Hindi. Therefore now both the target and source language has same synset identifier. When the word net has repeated occurrences from multi languages, it is referred to be Multidict. Each Multidict contains a row and column where as a row contains the words with the similar meaning and the column contains the synset identifier for the identified similar meanings.

SENSE BASED FEATURES

The sense based features[2] helps in the better understandability of the document and improves the accuracy in the Sentiment Analysis of any document. Implementing sense based features helps to overcome some of the issues such as,

- Evaluating the benefits and accuracy of the sense based features over the word based features.
- The problem of unrecognized words in the training dataset can be avoided by using the similarity metrics

Sense based features helps in generating the sentiment analysis classifier in a more superior way to avoid the severity of unidentified words in the testing dataset by replacing them with the identified words from the training dataset. This feature was not possible while using the word based representation as it does not support the comparison of vast data to check its similarity. POS based analysis has been made to improve the accuracy of the Sentiment Analysis. The comparison has been made between the samples created by the word based feature and the sense based feature. The synset replacement algorithm is used to replace the unidentified synset in the testing data set with the most relevant matching synset from the training data set. This helps in the improvement of the Performance of Sentiment Analysis. Some of the examples have been illustrated to show how the Sentiment Analysis works based on sense based features.

Sentence 1: “ Rama’s face fell when he heard that he has failed in the examination.”

Sentence 2: “ An Orange fell from the tree.”

In the above two sentences, the word "fell" appears in two different contexts. In sentence 1, the word "fell" represents the expression of sadness or the disappointment. In sentence 2, "fell" represents that the fruit has fallen due to the gravity of the Earth. While reading the sentence 1, the user opinion towards the sentence shows a negative polarity due to negative sense in the word "fell", whereas the sentence 2 does not infer any polarity, it is neither a positive opinion nor a negative opinion, therefore the sentence 2 does not infer any sentiment. By analyzing the second scenario, the word sense feature is more clearly understood.

Sentence 3: " King Cobra is deadly poisonous."

Sentence 4: " Rane spins the ball in a deadly way."

The word "deadly" in the above two sentences gives both positive as well as negative polarity. Sentence 3 infers a negative polarity as it is representing a dangerous entity, sentence 4 infers a positive polarity. Finally consider the third scenario,

Sentence 5: " Tip's language is very vulgar."

Sentence 6: " He behaves in a crude way."

In sentence 5 and 6, the words "vulgar" and "crude" are semantically similar in nature. These words can be identified with the synonymous nature by looking into its senses of the word. After analyzing all the three scenarios, one can conclude that the scenario 1 can be analyzed with the a sentiment and also a non-sentiment sentence. Scenario 2 has opposite polarities where as one infers positive sentiment and the other infers the negative sentiment. The third scenario demonstrates the senses of words that is the identification of the words with similar meaning in a synset.

2.1 Representation of the sense based words

By considering that both Hindi and Marathi have the same synset identifier, the words from each language has been represented with both the languages with the help of the corresponding identifier. For the target language, the training word set and testing word set is mapped for the respective synset identifier during the cross lingual settings[3]. Training dataset builds a classification model and tests it on testing data set. Since both wordset contains synset identifier, the experiment is classified into manually annotated dataset and the automatically annotated data set. Therefore the evaluation have been classified as manually annotated words and automatically annotated words generated by Word Sense Disambiguation Engine.

2.1.1 Sentiment classification using Naïve translation

A classifier developed using the training language is used to convert the testing language dataset to the training language dataset. Now the converted words are mapped from testing documents to the respective training dataset, Naïve translation is obtained as a result of this conversion. Semantic transfer or the syntactic transfer is not maintained. Multidict is used for the translation and two approaches are used for the replacement method i.e, Exact Word Replacement and the Random Word Replacement [4].

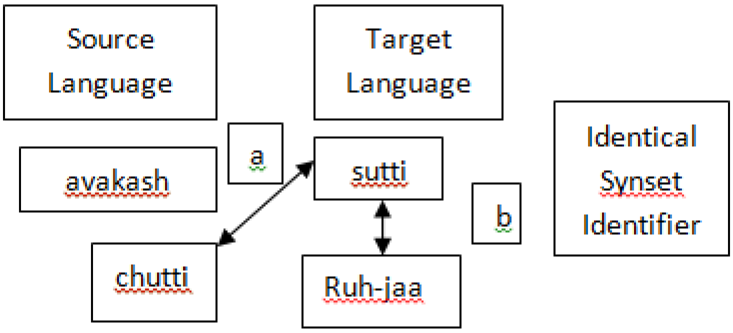

Fig-2: Example of Exact Word Replacement (a) and Random Word Replacement (b)

EXACT WORD REPLACEMENT

The word that is most similar is chosen and the respective word from source language is chosen and the replacement is done on a basis of disambiguation word sense identifier for the target language. In the figure 2, for the target word “sutti”, similar word from the source “chutti” is selected (a).

Random Word Replacement: Some random word having the identical sysnset identifier is replaced. For example, in the figure 2, the random word “Ruh-jaa” is chosen for the word “sutti”.

The dataset for Hindi and Marathi has been collected from the tourism review document which has been assigned with a positive and negative polarity by a native reviewer. Hindi and Marathi corpora consist of 11038 and 12566 words respectively. Each review consists of four to five sentences of approximately 20 words. A native reviewer has made the Sentiment Analysis on this tourist document and has provided 100 positive and negative reviews for the Hindi document and 75 positive and negative reviews for the Marathi document. The native speaker has manually annotated the words which are used as the manually annotated dataset. On the basis of the POS tagging, annotation tool assigns each word to all the possible sense entries. Then the right matching word is chosen by the lexicographer based on the context.

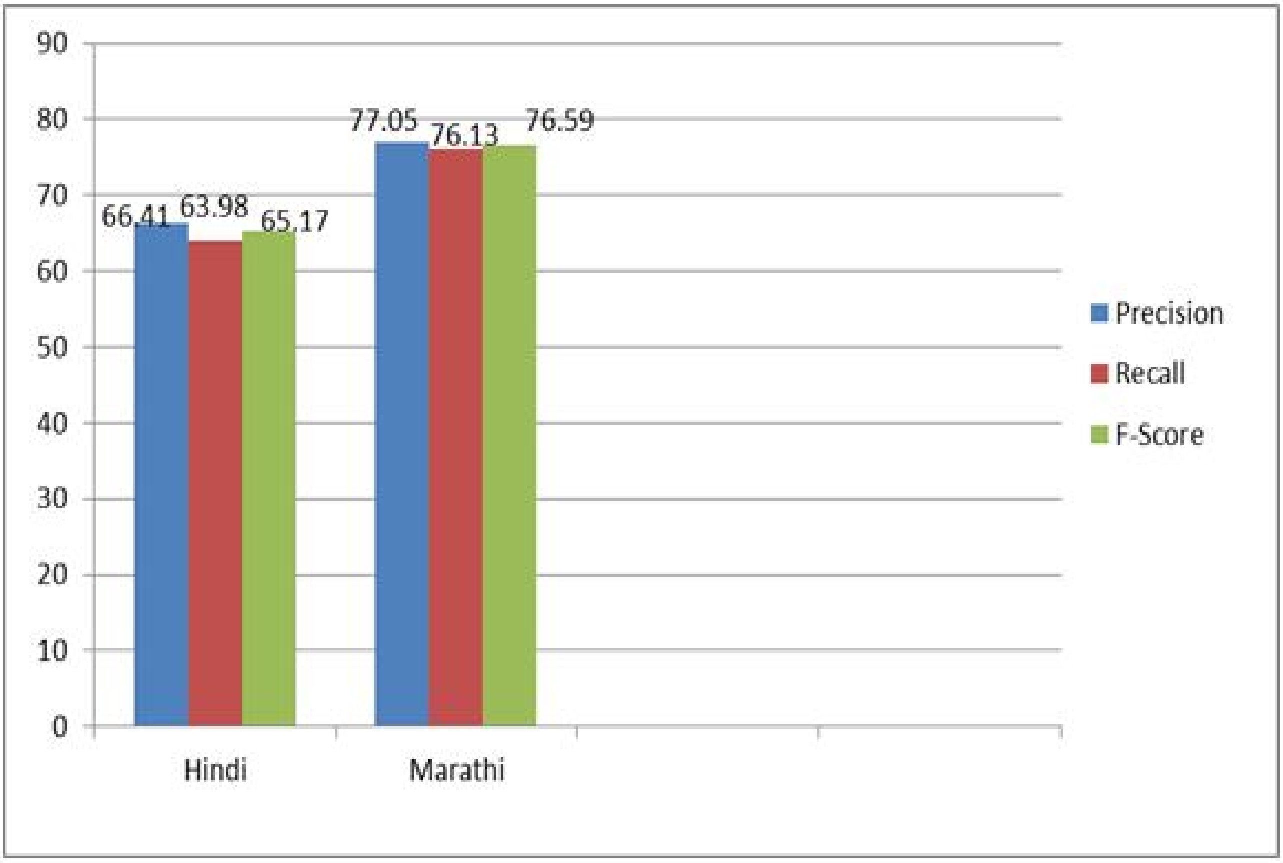

Figure-3: Annotation statistics for Hindi and Marathi.

Figure 3 represents the overall precision, recall and f-score for Hindi and Marathi dataset. The p precision, recall and f-score has been calculated separately for noun, adverb, verb and adjective and a overall average has been provided for both Hindi and Marathi dataset.

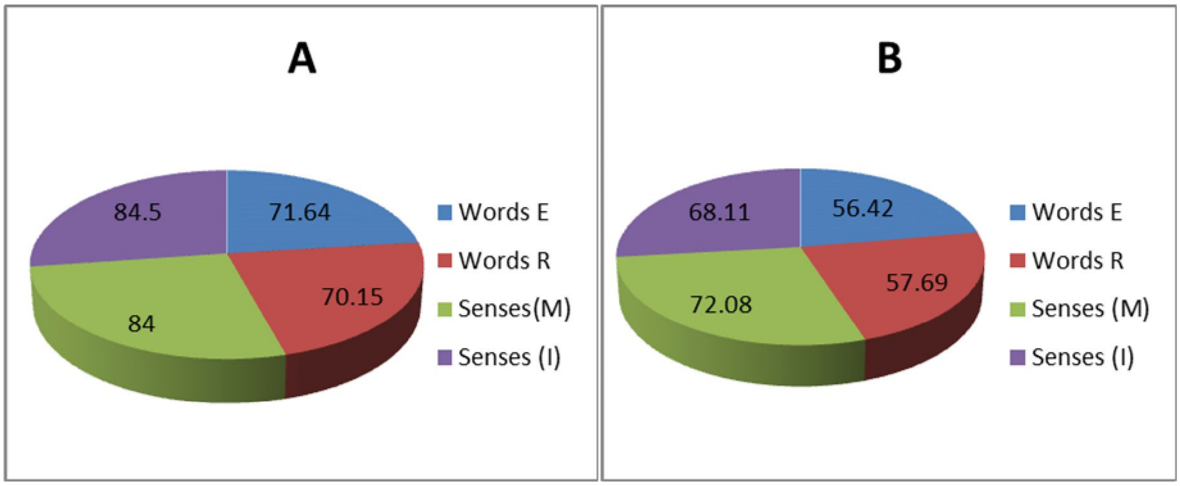

Figure-4: Cross Lingual Sentiment Analysis Accuracy when Hindi is used as training language (A) and Marathi is used as training language (B)

Figure 4 represents the accuracy of CLSA when Hindi is used as training language and Marathi is used as the testing language in the pie-chart (A), and when Hindi is used as testing language and Marathi is used as the training language in the pie-chart (B). The exact match word is represented by Words E and randomly selected words are represented by Words R. Senses (M) is where the word senses are used for Marathi and Senses (I) where the word senses are used for Hindi. When Hindi is used as the target language, some improvements can be seen in the positive recall than using Marathi as a target language. This approach can be used as an alteration method for the Machine Translation method based CLSA approach since most of the Indian languages do not use the Machine Translation method. Hindi word net consists of finer sense details in its word net than Marathi, therefore Hindi have a less accuracy while compared to Marathi. This may result in assigning an improper sense for a word in Hindi more than that of the chances in Marathi. Also when Hindi is used as a testing language, the CLSA accuracy falls down. This low accuracy is also because of the reason that Marathi has a very less corpora containing only few training samples.

Due to the missing concepts and the defect in the Hindi morphology analyzer, the usage of this approach may fall back. Since Marathi is expanded by taking the words from Hindi, most of the concepts that are present in Marathi are derived from Hindi; some of the concepts which are present in Hindi may not be included in the Marathi word net which leads low accuracy in sense based CLSA. Morphology analyzer searches for the deep words for the verbs which do not match the words present in Hindi word net and this causes the lower accuracy for Hindi.

2.2 Analysis with POS tagging

With the analysis of POS tagging, two comparisons has been made,

- Trained words associated with the particular POS
- Trained word senses with the particular POS

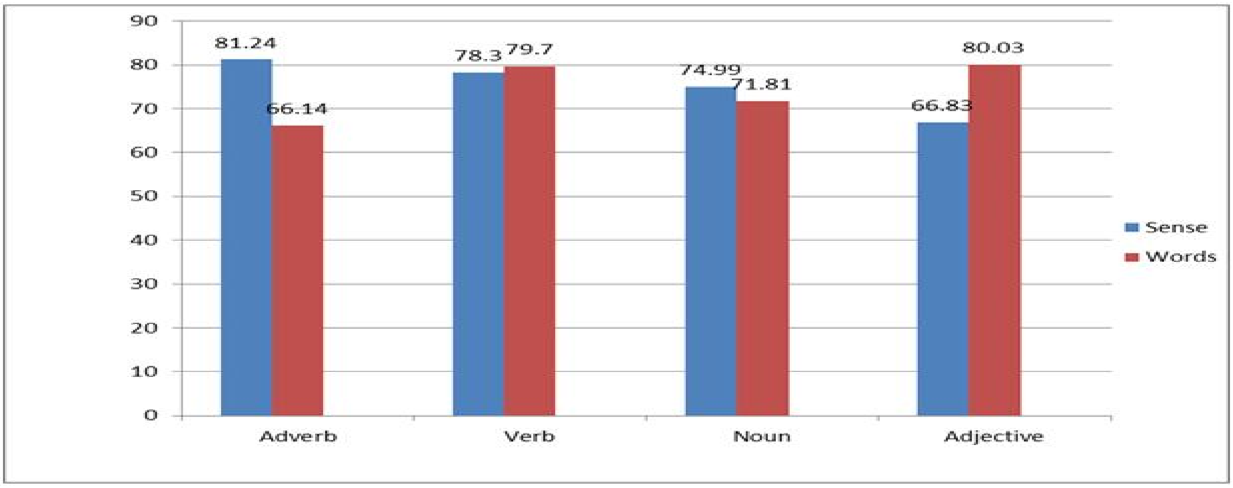

Fig05: Manually annotated data set using POS tagging with x-axis representing POS category and y-axis representing the accuracy

Figure 5 represents the accuracy that has been obtained with the classification of the POS tagging for the manual word senses and just words. The performance of the classification based on the POS is directly affected by the adjectives in the lexeme space [5]. The performance of Sentiment Analysis has been more affected with the disambiguation of verb and adverb synsets than the disambiguation of adjectives and noun. Sentiment Analysis is directly conveyed with the use of the adjectives but this direct use of adjectives alone without using any other word sense features may result in the low accuracy. At some point of time, the sentiments may be difficult to analyze and it may be difficult to express directly by using adjectives.

Due to some drawbacks in the word sense features such as, low disambiguation accuracy, negation handling, interjections word net coverage and document specificity, word sense features could not sustain for a longer time.

Low disambiguation accuracy: It all depends on the annotation system used for the Sentiment Analysis of automatic word sense annotation.

Negation handling: While handling the negation entity, the words are considered as units for Sentiment Analysis. Since syntax is neglected in understanding the sentiment, the positive sentences containing the negative words is considered to be the negative polarity sentences. For example, “She cried in the celebration of her happiness." Even though the sentence represents her happiness, because of the word "cried", it is inferring a negative polarity.

Interjections word net coverage: Some of the interjections are not included in the word net and hence it fails to get disambiguated. The words like "wow", "yeahh" are not included in the word net and cannot be given any polarity.

Document specificity: Since one document contains the description of many concepts, it is difficult to make a overall Sentiment Analysis. For example the document of trip may be containing the information regarding to the public behavior of that country, its tradition, traffic in the respective cities, tourist places and hence it may contain difference of opinion which fails to assign a overall Sentiment to the document [6].

3. BILINGUALLY CONSTRAINED RECURSIVE AUTO-ENCODER (BRAE)

This approach aims to learn the vector representation for both source and the target language. Before getting into BRAE, bilingual document representation learning (BiDRL) [7] is important. The architecture is shown in figure 6. Both monolingual and bilingual constraints are proposed to learn the model after obtaining the dataset from source and target language. Words are built in each individual language with the help of monolingual model. The consistent embedding space is built between two languages with the help of bilingual model. Joint learning is a semi supervised model and it uses the sentiment labels from the training dataset. Wit the use of this approach, both semantic and sentiment relationship can be achieved. The existing algorithms make use of semantic connection only where as the BRAE approach learns the representation of word and document together. The previous algorithm just provides the average of the total words present in the dataset but the BRAE provides high embedding performance with the use of sentiment labels which helps in creating a bridge between two languages. In monolingual model, both the words and documents are mapped to unique vectors in the paragraph vector. Each document is considered to be a unique token with respect to the context of entire word set present in the document. This makes every word in the document to be predicted easily by the tokens.

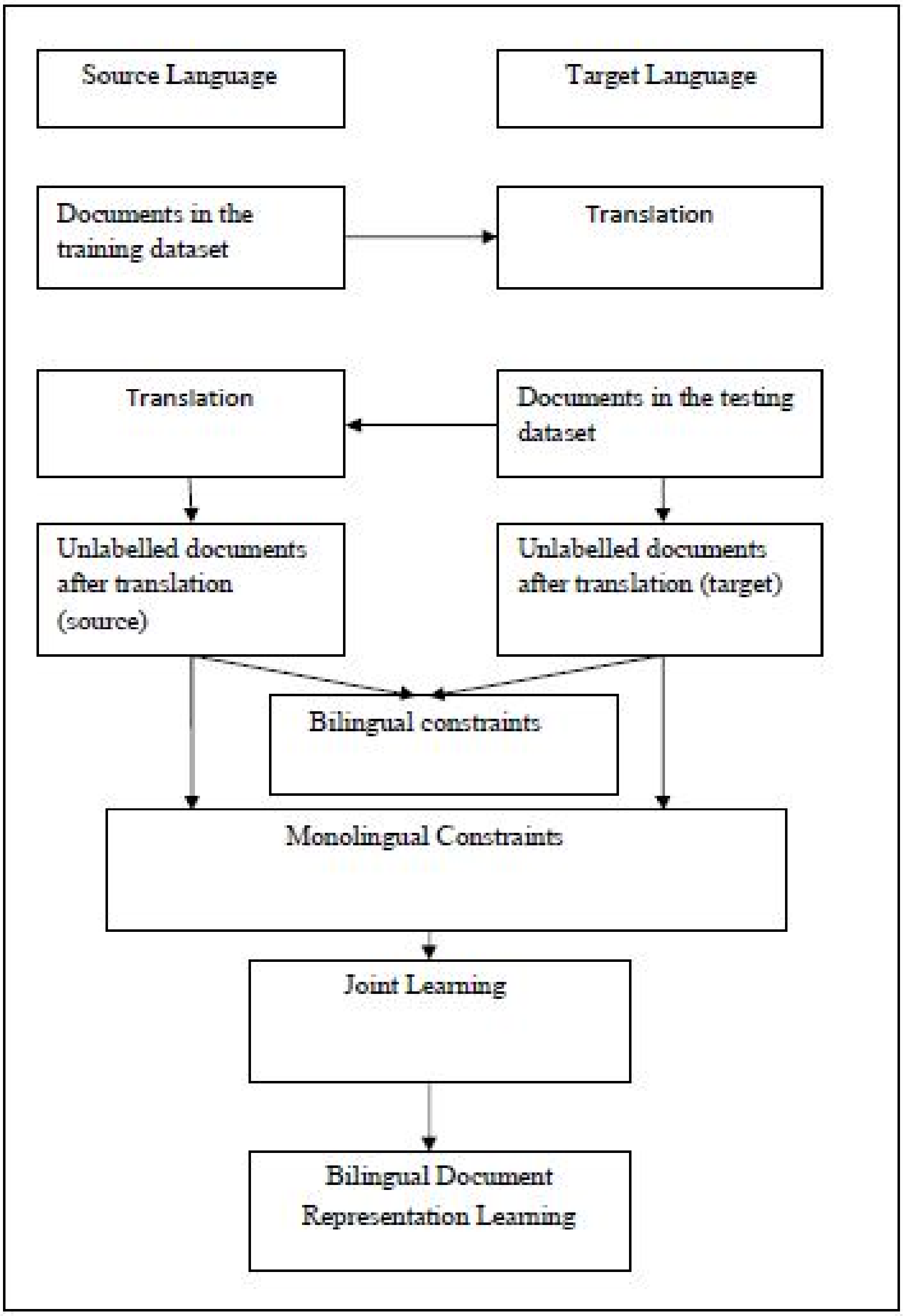
Source Language
Target Language
Documents in the training dataset
Translation
Translation
Documents in the testing dataset
Unlabelled documents after translation (source)
Unlabelled documents after translation (target)
Bilingual constraints
Monolingual Constraints
Joint Learning
Bilingual Document Representation Learning

Figure-6: Architecture of BiDRL

Drawback of the bilingual model learning is that embedding space over source and target language must be consistent. Therefore three strategies are proposed to overcome the language gap. Introducing the logistic regressions is the first approach which helps the documents with same sentiment to categorize under same area in the embedding space. The second approach is to reduce the distance between original and the translated document. Final approach is to generate the similar representation of text for the same sentiment [8].

UNSUPERVISED TRAINING PHASE OF BRAE

Recursive Auto Encoder training and the cross training are the two phases in unsupervised training of BRAE. In RAE training, RAE framework has to be applied and source and target phrases have to be pre-trained. In cross training phase, the source phrase will be updated using the target phrase representation and source phrase representation will be obtained. And also target phrase will be updated using the source phrase representation and target phrase representation will be obtained. This iteration will be repeating until the cross joint error is minimized and the phase is terminated [9].

SUPERVISED TRAINING PHASE OF BRAE

Each source and the target language is trained using the labeled monolingual dataset. The first phase in supervised training is training for the resource rich language that is for the source language. Here the parameters of RAE source language are modified. The next phase is training of the resource poor language or the target language. Here the parameters of RAE target language are modified and the final phase is to predict the overall sentiment of the document.

Overall sentiment is predicted in association with the target language using the phrase embeddings of the top layer [10].

CONCLUSION

The major advantage of this BRAE model is because of the usage of the words that are not present in the labeled data set in the test data. The model was successful in assigning a polarity even for the unknown word by comparing it with the word which is semantically similar in nature. The model provided a benefit of obtaining exact grammatical phrases and it was successful in inferring the correct polarity to the polysemy words. This eliminates the need of manually annotated dataset and the word net. For example, the sentence "She cried out with her daughter fame in the media" assigns the positive polarity with BRAE model where as other baseline models assigned the negative polarity. Some errors were found where the model actually failed to assign the correct sentiment for the sentence like "He was destroying the fun every time, but this time he dint do it". This sentence was assigned with negative polarity even though it represents the positivity. And also the sentence "This performance made her previous to look good" assigned a positive polarity instead of assigning a negative one. The model failed in subtle contextual sentences. The future enhancements that can be made to the BRAE model are to learn the phrase embeddings for multiple languages at the same time. Paraphrase detection can be done by applying the BRAE framework to this cross lingual tasks.

REFERENCES

1. Bilingually-constrained Phrase Embeddings for Machine Translation Jiajun Zhang, Shujie Liu, Mu Li, Ming Zhou, Chengqing Zong, National Laboratory of Pattern Recognition, CASIA, Beijing, P.R. China

2. Cross-Lingual Sentiment Analysis for Indian Languages using Linked Word Nets Balamurali A R1,2 Adit ya Joshi1 Pushpak Bhattachar y ya1, Proceedings of COLING 2012: Posters, pages 73–82,COLING 2012, Mumbai, December 2012.

3. Peng Li, Yang Liu, Maosong Sun. 2013. Recursive autoencoders for itg-based translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing.

4. Harnessing WordNet Senses for Supervised Sentiment Classification Balamurali A R1,2 Aditya Joshi2 Pushpak Bhattacharyya2 Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 1081–1091, Edinburgh, Scotland, UK, July 27–31, 2011.c 2011 Association for Computational Linguistics

5. [5] Cross-Lingual Sentiment Classification with Bilingual Document Representation Learning Xinjie Zhou, Xianjun Wan and Jianguo Xiao, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1403–1412, Berlin, Germany, August 7-12, 2016.c 2016 Association for Computational Linguistics

6. Jianfeng Gao, Xiaodong He, Wen-tau Yih, and Li Deng. 2013. Learning semantic representations for the phrase translation model. arXiv preprint arXiv:1312.0482.

7. Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang. 2013. Decoding with largescale neural language models improves translation. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1387–1392.

8. Nan Yang, Shujie Liu, Mu Li, Ming Zhou, and Nenghai Yu. 2013. Word alignment modeling with context dependent deep neural network. In 51st Annual Meeting of the Association for Computational Linguistics.

9. Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1393–1398.

10. Richard Zens, Daisy Stanton, and Peng Xu. 2012. A systematic comparison of phrase table pruning techniques. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 972–983