The Haves and the Have-Nots: Leveraging Unlabelled Corpora for Sentiment Analysis

Kashyap Popat[2] **Balamurali A R**[1,2,3] **Pushpak Bhattacharyya**[2] **Gholamreza Haffari**[3]
[1]IITB-Monash Research Academy, IIT Bombay
[2]Dept. of Computer Science and Engineering, IIT Bombay
{kashyap,balamurali,pb}@cse.iitb.ac.in
[3]Monash University
Australia
reza@monash.edu

Abstract

Expensive feature engineering based on WordNet senses has been shown to be useful for document level sentiment classification. A plausible reason for such a performance improvement is the reduction in data sparsity. However, such a reduction could be achieved with a lesser effort through the means of syntagma based word clustering. In this paper, the problem of data sparsity in sentiment analysis, both monolingual and cross-lingual, is addressed through the means of clustering. Experiments show that cluster based data sparsity reduction leads to performance better than sense based classification for sentiment analysis at document level. Similar idea is applied to Cross Lingual Sentiment Analysis (CLSA), and it is shown that reduction in data sparsity (after translation or bilingual-mapping) produces accuracy higher than Machine Translation based CLSA and sense based CLSA.

1 Introduction

Data sparsity is the bane of Natural Language Processing (NLP) (Xue et al., 2005; Minkov et al., 2007). Language units encountered in the test data but absent in the training data severely degrade the performance of an NLP task. NLP applications innovatively handle data sparsity through various means. A special, but very common kind of data sparsity *viz.*, word sparsity, can be addressed in one of the two obvious ways: 1) sparsity reduction through *paradigmatically related* words or 2) sparsity reduction through *syntagmatically related* words.

Paradigmatic analysis of text is the analysis of concepts embedded in the text (Cruse, 1986; Chandler, 2012). WordNet is a byproduct of such an analysis. In WordNet, paradigms are manually generated based on the principles of lexical and semantic relationship among words (Fellbaum, 1998). WordNets are primarily used to address the problem of word sense disambiguation. However, at present there are many NLP applications which use WordNet. One such application is Sentiment Analysis (SA) (Pang and Lee, 2002). Recent research has shown that word sense based semantic features can improve the performance of SA systems (Rentoumi et al., 2009; Tamara et al., 2010; Balamurali et al., 2011) compared to word based features.

Syntagmatic analysis of text concentrates on the surface properties of the text. Compared to paradigmatic property extraction, syntagmatic processing is relatively light weight. One of the obvious syntagmas is *words*, and words are grouped into equivalence classes or clusters, thus reducing the model parameters of a statistical NLP system (Brown et al., 1992). When used as an additional feature with word based language models, it has been shown to improve the system performance *viz.*, machine translation (Uszkoreit and Brants, 2008; Stymne, 2012), speech recognition (Martin et al., 1995; Samuelsson and Reichl, 1999), dependency parsing (Koo et al., 2008; Haffari et al., 2011; Zhang and Nivre, 2011; Tratz and Hovy, 2011) and NER (Miller et al., 2004; Faruqui and Padó, 2010; Turian et al., 2010; Täckström et al., 2012).

In this paper, the focus is on alleviating the data sparsity faced by supervised approaches for SA through the means of cluster based features. As WordNets are essentially word

clusters wherein words with the same meaning are clubbed together, they address the problem of data sparsity at word level. The abstraction and dimensionality reduction thus achieved attributes to the superior performance for SA systems that employs WordNet senses as features. However, WordNets are manually created. Automatic creation of the same is challenging and not much successful because of the linguistic complexity involved. In case of SA, manually creating the features based on WordNet senses is a tedious and an expensive process. Moreover, WordNets are not present for many languages. All these factors make the paradigmatic property based cluster features like WordNet senses a less promising pursuit for SA.

The syntagmatic analysis essentially makes use of distributional similarity and may in many circumstances subsume the paradigmatic analysis. In the current work, this particular insight is used to solve the data sparsity problem in the sentiment analysis by leveraging unlabelled monolingual corpora. Specifically, experiments are performed *to investigate whether features developed from manually crafted clusterings (coming from WordNet) can be replaced by those generated from clustering based on syntagmatic properties*.

Further, *cluster based features are used to address the problem of scarcity of sentiment annotated data in a language*. Popular approaches for Cross-Lingual Sentiment Analysis (CLSA) (Wan, 2009; Duh et al., 2011) depend on Machine Translation (MT) for converting the labeled data from one language to the other (Hiroshi et al., 2004; Banea et al., 2008; Wan, 2009). However, many languages which are truly resource scarce, do not have an MT system or existing MT systems are not ripe to be used for CLSA (Balamurali et al., 2013). To perform CLSA, this study leverages unlabelled parallel corpus to generate the word alignments. These word alignments are then used to link cluster based features to obliterate the language gap for performing SA. No MT systems or bilingual dictionaries are used for this study. Instead, language gap for performing CLSA is bridged using linked cluster or *cross-lingual* clusters (explained in section 4) with the help of unlabelled monolingual corpora. The contributions of this paper are two fold:

1. *Features created from manually built and finer clusters can be replaced by inexpensive cluster based features generated solely from unlabelled corpora.* Experiments performed on four publicly available datasets in three languages *viz., English, Hindi* and *Marathi*[1] suggest that cluster based features can considerably boost the performance of an SA system. Moreover, state of the art result is obtained for one of the publicly available dataset.

2. *An alternative and effective approach for CLSA is demonstrated using clusters as features*. Word clustering is a powerful mechanism to "transfer" a sentiment classifier from one language to another. Thus can be used in truly resource scarce scenarios like that of *English-Marathi* CLSA.

The rest of the paper is organized as follows: section 2 presents related work. Section 3 explains different word cluster based features employed to reduce data sparsity for monolingual SA. In section 4, alternative CLSA approaches based on word clustering are elucidated. Experimental details are explained in section 5. Results and discussions are presented in section 6 and section 7 respectively. Finally, section 8 concludes the paper pointing to some future research possibilities.

2 Related Work

The problem of SA at document level is defined as the classification of document into different polarity classes (positive and negative) (Turney, 2002). Both supervised (Benamara et al., 2007; Martineau and Finin, 2009) and unsupervised approaches (Mei et al., 2007; Lin and He, 2009) exist for this task.

Supervised approaches are popular because of their superior classification accuracy (Mullen and Collier, 2004; Pang and Lee, 2008). Feature engineering plays an important role in these systems. Apart from the commonly used bag-of-words features based on unigrams/bigrams/ngrams (Dave et al., 2003; Ng et al., 2006; Martineau and Finin, 2009),

[1]Hindi and Marathi belong to the Indo-Aryan subgroup of the Indo-European language family and are two widely spoken Indian languages with a speaker population of 450 million and 72 million respectively.

syntax (Matsumoto et al., 2005; Nakagawa et al., 2010), semantic (Balamurali et al., 2011) and negation (Ikeda et al., 2008) have also been explored for this task. There has been research related to clustering and sentiment analysis. In Rooney et al. (2011), documents are clustered based on the context of each document and sentiment labels are attached at the cluster level. Zhai et al. (2011) attempts to cluster features of a product to perform sentiment analysis on product reviews. In this work, word clusters (syntagmatic and paradigmatic) encoding a mixture of syntactic and semantic information are used for feature engineering.

In situations where labeled data is not present in a language, approaches based on cross-lingual sentiment analysis are used. Most often these methods depend on an intermediary machine translation system (Wan, 2009; 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 the sentiment can be expressed which itself manifested as a result of cultural diversity amongst different languages, an MT system has to be of a superior quality to capture them.

3 Clustering for Sentiment Analysis

The goal of this paper, to remind the reader, is to investigate whether superior word cluster features based on manually crafted and fine grained lexical resource like WordNet can be replaced with the syntagmatic property based word clusters created from unlabelled monolingual corpora.

In this section, different clustering approaches are presented for feature engineering in a monolingual setting.

3.1 Approach 1: Clustering based on WordNet Sense

A synonymous set of words in a WordNet is called a synset. Each synset can be considered as a word cluster comprising of semantically similar words. Balamurali et al. (2011) showed that WordNet synsets can act as good features for document level sentiment classification.

Motivation for their study stems from the fact that different senses of a word can have different polarities. To empirically prove the superiority of sense based features, different variants of a travel review domain corpus were generated by using automatic/manual sense disambiguation techniques. Thereafter, accuracies of classifiers based on different sense-based and word-based features were compared. The results suggested that WordNet synset based features performed better than word-based features.

In this study, synset identifiers are extracted from manually/automatically sense annotated corpora and used as features for creating sentiment classifiers. The classifier thus build is used as a baseline. Apart from this, another baseline employing word based features are used for a comprehensive comparison.

3.2 Approach 2: Syntagmatic Property based Clustering

For this particular study, a co-occurrence based algorithm is used to create word clusters. As the algorithm is based on co-occurrence, one can extract the classes that have the flavour of syntagmatic grouping, depending on the nature of underlying statistics. Agglomerative clustering algorithm by Brown et al. (1992) is used for this purpose. It is a hard clustering algorithm *i.e.*, each word belongs to one cluster only.

Formally, as mentioned in Brown et al. (1992), let C be a hard clustering function which maps vocabulary V to one of the K clusters. Then, the likelihood ($L()$) of a sequence of word tokens, $w = [w_j]_{j=1}^{m}$, with $w_j \in V$, can be factored as,

$$L(w;C) = \prod_{j=1}^{m} p(w_j|C(w_j))p(C(w_j)|C(w_{j-1}))) \quad (1)$$

Words are assigned to clusters such that the above quantity is maximized. For the purpose of sentiment classification, cluster identifiers representing words in the document are used as features for training.

4 Clustering for Cross Lingual Sentiment Analysis

Existing approaches for CLSA depend on an intermediary machine translation system to bridge the language gap (Hiroshi et al., 2004; Banea et al., 2008). Machine translation is very resource intensive. If a language is truly resource scarce, it is mostly unlikely to have an MT system. Given that sentiment analysis is a less resource intensive task compared to machine translation, the use of an MT system is hard to justify for performing

CLSA. As a viable alternative, cluster linkages could be learned from a bilingual parallel corpus and these *linkages* can be used to bridge the language gap for CLSA.

In this section, three approaches using clusters as features for CLSA are compared. The language whose annotated data is used for training is called the source language (S), while the language whose documents are to be sentiment classified is referred to as the target language (T).

4.1 Approach 1: Projection based on Sense (PS)

In this approach, a Multidict is used to bridge the language gap for SA. A Multidict is an instance of WordNet where the same sense from different languages are linked (Mohanty et al., 2008). An entry in the multidict will have a WordNet sense identifier from S and the corresponding WordNet sense identifier from T. The approach of projection based on sense is explained in Algorithm 1. Note that after the *Sense Mark* operation, each document will be represented as a vector of WordNet sense identifiers.

Algorithm 1 Projection based on sense

Input: Polarity labeled data in source language (S) and data in target language (T) to be labeled

Output: Classified documents

1: Sense mark the polarity labeled data from S
2: Project the sense marked corpora from S to T using a Multidict
3: Model the sentiment classifier using the data obtained in step-2
4: Sense mark the unlabelled data from T
5: Test the sentiment classifier on data obtained in step-4 using model obtained in step-3

Sense identifiers are the features for the classifier. For those sense identifiers which do not have a corresponding entry in the Multidict, no projection is performed.

4.2 Approach 2: Direct Cluster Linking (DCL)

Given a parallel bilingual corpus, word clusters in S can be aligned to clusters in T. Word alignments are created using parallel corpora. Given two aligned word sequences $w^S = [w_j^S]_{j=1}^m$ and $w^T = [w_k^T]_{k=1}^n$, let $\alpha^{T|S}$ be a set of scored alignments from the source language to the target language. Here, an alignment from the $a_k{}^{th}$ source word to the k^{th} target word, with score $s_{k,a_k} > \varepsilon$ is represented as $(w_k^T,\ w_{a_k}^S,\ s_{k,a_k}) \in \alpha^{T|S}$. To simplify, $k \in \alpha^{T|S}$ is used to denote those target words w_k^T that are aligned to some source word $w_{a_k}^S$.

The source and the target side clusters are linked using the Equation (2).

$$LC(l) = \underset{t}{\operatorname{argmax}} \sum_{\substack{k \in \alpha^{T|S} \cup \alpha^{S|T} \\ s.t. C^T(w_k^T)=t \\ C^S(w_{a_k}^S)=l}} s_{k,a_k} \qquad (2)$$

Here, a target side cluster $t \in C^T$ is linked to a source side cluster $l \in C^S$ such that the total alignment score between words in l and words in t is maximum. C^S and C^T stands for source and target side cluster list respectively. $LC(l)$ gives the target side cluster t to which l is linked.

4.3 Approach 3: Cross-Lingual Clustering (XC)

Direct cluster linking approach suffers from the size of alignment dataset in the form of parallel corpora. The size of the alignment dataset is typically smaller than the monolingual dataset. To circumvent this problem, Täckström et al. (2012) introduced cross-lingual clustering. In cross-lingual clustering, the objective function maximizes the joint likelihood of monolingual and cross-lingual factors. Given a list of words and clusters it belongs to, a clustering algorithm tries to obtain word-cluster association which maximizes the joint likelihood of words and clusters. Whereas in case of cross-lingual clustering, the same clustering can be explained in terms of maximizing the likelihood of monolingual word-cluster pairs of the source, the target and alignments between them.

Formally, as stated in Täckström et al. (2012), Using the model of Uszkoreit and Brants (2008), the likelihood of a sequence of word tokens, $w = [w_j]_{j=1}^m$, with $w_j \in V$, can be factored as,

$$L(w; C) = \prod_{j=1}^{m} p(w_j|C(w_j))p(C(w_j)|w_{j-1})) \qquad (3)$$

Note this is different from the likelihood estimation of Brown et al. (1992) (Equation (1)), where $C(w_j)$ was conditioned on $C(w_{j-1})$. This

makes the computation easier as suggested in the original paper. The Equation (3) in a cross lingual setting will be transformed as given below:

$$L^{S,T}(w^S, w^T; \alpha^{T|S}, \alpha^{S|T}, C^S, C^T) = L^S(...).L^T(...).L^{T|S}(...).L^{S|T}(...) \quad (4)$$

Here, $L^{T|S}(...)$ and $L^{S|T}(...)$ are factors based on word alignments, which can be represented as:

$$L^{T|S}(w^T; \alpha^{T|S}, C^T, C^S) = \prod_{k \in \alpha^{T|S}} p(w_k^T | C^T(w_k^T)) p(C^T(w_k^T) | C^S(w_{a_k}^S))) \quad (5)$$

Based on the optimization objective in Equation (4), a pseudo algorithm is defined in Algorithm 2. For more information, readers are requested to refer Täckström et al. (2012).

Algorithm 2 Cross-lingual Clustering (XC)

Input: Source and target language corpus
Output: Cross-lingual clusters
1: ## C^S, C^T *randomly initialized*
2: **for** $i \leftarrow 1$ to N **do**
3: Find $C_*^S \approx \text{argmax}_{C^S} L^S(w^S; C^S)$
4: Project C_*^S to C^T
5: Find $C_*^T \approx \text{argmax}_{C^T} L^T(w^T; C^T)$
6: Project C_*^T to C^S
7: **end for**

An MT based CLSA approach is used as the baseline. Training data from S is translated to T and classification model is learned using unigram based features. Thereafter, the classifier is directly tested on data from T.

5 Experimental Setup

Analysis was performed on three languages, *viz.*, *English (En), Hindi (Hi) and Marathi (Mar)*. CLSA was performed on two language pairs, *English-Hindi* and *English-Marathi*. For clustering the words, monolingual data of Indian Languages Corpora Initiative (ILCI)[2] was used. It should also be noted that sentiment annotated data was also included in the data used for the word clusterings process. For Brown clustering, an implementation by Liang (2005) was used. Cross-lingual clustering for CLSA was implemented as directed in Täckström et al. (2012).

Monolingual SA: For experiments in *English*, two polarity datasets were used. The first one (*En-TD*) by Ye et al. (2009) contains user-written reviews on travel destinations. The dataset consists of approximately 600 positive and 591 negative reviews. Reviews were also manually sense annotated using WordNet 2.1. The sense annotation was performed by two annotators with an inter-annotation agreement of 93%. The second dataset (*En-PD*)[3] on product reviews (music instruments) from Amazon by Blitzer et al. (2007) contains 1000 positive and 1000 negative reviews. This dataset was sense annotated using an automatic WSD engine which was trained on tourism domain (Khapra et al., 2010). Experiments using this dataset were done to study the effect of domain on CLSA. For experiments in *Hindi* and *Marathi*, polarity datasets by Balamurali et al. (2012) were used.[4] These are reviews collected from various *Hindi* and *Marathi* blogs and Sunday editorials. *Hindi* dataset consist of 98 positive and 100 negative reviews. Whereas *Marathi* dataset contains 75 positive and 75 negative reviews. Apart from being marked with polarity labels at document level, they are also manually sense annotated using *Hindi* and *Marathi* WordNet respectively.

CLSA: The same datasets used in SA are also used for CLSA. Three approaches (as described in section 4) were tested for *English-Hindi* and *English-Marathi* language pairs. To create alignments, *English-Hindi* and *English-Marathi* parallel corpora from ILCI were used. *English-Hindi* parallel corpus contains 45992 sentences and *English-Marathi* parallel corpus contains 47881 sentences. To create alignments, GIZA++[5] was used (Och and Ney, 2003).

As a preprocessing step, all stop words were removed. Stemming was performed on English and Hindi whereas for Marathi data, Morphological Analyzer was used to reduce the words to their respective lemmas.

All experiments were performed using C-SVM

[2] http://sanskrit.jnu.ac.in/ilci/index.jsp

[3] http://www.cs.jhu.edu/~mdredze/datasets/sentiment/

[4] http://www.cfilt.iitb.ac.in/resources/senti/MPLC_tour_downloaderInfo.php

[5] http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html

Features	En-TD	En-PD	Hi	Mar
Words	87.02	77.60	77.36	92.28
WordNet Sense (Paradigmatic)	89.13	74.50	**85.80**	96.88
Clusters (Syntagmatic)	**97.45**	**87.80**	83.50 ✠	**98.66**

Table 1: Classification accuracy for monolingual sentiment analysis. For English, results are reported on two publicly available datasets based on Travel Domain (TD) and Product Domain (PD).

Features	Words	Clust-200	Clust-500	Clust-1000	Clust-1500	Clust-2000	Clust-2500	Clust-3000
En-TD	87.02	97.37	97.45	96.94	96.94	96.52	96.52	96.52
En-PD	77.60	73.20	82.30	84.30	86.35	86.45	87.80	87.40

Table 2: Classification accuracy (in %) versus cluster size (number of clusters to be used).

(linear kernel with parameter optimized over training set using 5 fold cross validation) available as a part of LibSVM package[6]. SVM was used since it is known to perform well for sentiment classification (Pang et al., 2002). Results reported are based on the average of ten-fold cross-validation accuracies. Standard text metrics are used for reporting the experimental results.

6 Results

Monolingual classification results are shown in Table[7]1. Table shows accuracies of SA systems developed on feature set based on words, senses and clusters. It must be noted that accuracies reported for cluster based features are with respect to the best accuracy based on different cluster sizes. The improvements in results of cluster features based approach is found to be statistically significant over the word features based approach and sense features based approach at 95% confidence level when tested using a paired t-test (except for *Hindi* cluster features based approach). But in general, their accuracies do not significantly vary after cluster size crosses 1500.

Table 2 shows the classification accuracy variation when cluster size is altered. For, En-TD and En-PD experiments, the cluster size was varied between 200-3000 with an interval of 500 (after a size of 500). In the En-TD experiment, the best accuracy is achieved for cluster size 500, which is lesser than the number of unique-words/unique-senses (6435/6004) present in the data. Similarly, for the En-PD experiment, the optimal cluster size of 2500 is also lesser than the number of unique-words/unique-senses (30468/4735) present in the data.

To see the effect of training data size variation for different SA approaches in the En-TD experiment, the training data size is varied between 50 to 500. For this, a test set consisting of 100 positive and 100 negative documents is fixed. The training data size is varied by selecting different number of documents from rest of the dataset ($\sim$500 negative and $\sim$500 positive) as a training set. For each training data set 10 repeats are performed, *e.g.*, for training data size of 50, 50 negative and 50 positive documents are randomly selected from the training data pool of $\sim$500 negative and $\sim$500 positive. This was repeated 10 times (with replacement). The results of this experiment are presented in Figure 1.

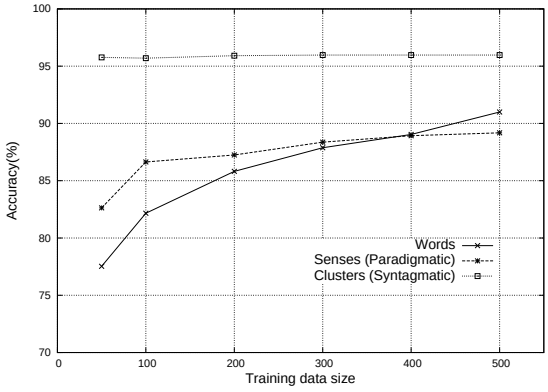

Figure 1: Training data variation on En-TD dataset.

Cross-lingual SA accuracies are presented in Table 3. As in monolingual case, the reported accuracies are for features based on the best cluster size.

[6] `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

[7] All results reported here are based on 10-fold except for Marathi (2-fold-5-repeats), as it had comparatively lesser data samples.

Target Language	MT	PS	DCL	XC
T=*Hi*	63.13	53.80	51.51	**66.16**
T=*Mar*	NA	54.00	56.00	**60.30**

Table 3: Cross-Lingual SA accuracy (%) on T=*Hi* and T=*Mar* with S=*En* for different approaches (**MT**=Machine Translation, **PS**=Projection based on Sense, **DCL**=Direct Cluster Linking , **XC**=Cross-Lingual Clustering. There is no MT system available for (S=*En*, T=*Mar*).

7 Discussions

In this section, some important observations from the results are discussed.

1. Syntagmatic analysis may be used in lieu of paradigmatic analysis for SA: The results suggest that word cluster based features using syntagmatic analysis is comparatively better than cluster (sense) based features using paradigmatic analysis. For two datasets in *English* and for the one in *Marathi* this holds true. For *English*, the gap between classification accuracy based on sense features and cluster features is around 10%. A state-of-art accuracy is obtained for the public dataset on travel domain (*En-TD*).

The difference in accuracy reduces as the language gets morphologically rich. In a morphologically rich language, morphology encompasses syntactical information, limiting the context it can provide for clustering. This can be seen from the classification results on *Marathi*. However for *Hindi*, classifier built on features based on syntagmatic analysis trails the one based on paradigmatic analysis.

Compared to *Marathi*, *Hindi* is a less morphologically rich language, hence, a better result was expected. However, a contrary result was obtained.✠ In *Hindi*, the subject and the object of the sentence are linked using a case marker. Upon error analysis, it was found that there was a lot of irregular compounding based on case markers. Case markers were compounded with the succeeding word. This is a deviation from the real scenario which would have resulted in incorrect clustering leading to an unexpected result. However, the same would not have occurred for a classifier developed on sense based features as it was manually sense tagged.

Clustering induces a reduction in the data sparsity. For example, on *En-PD*, percentage of features present in the test set and not present in the training set to those present in the test set are *34.17%, 11.24%, 0.31%* for words, synsets and cluster based features respectively. The improvement in the performance of classifiers may be attributed to this feature size reduction. However, it must be noted that clustering based on unlabelled corpora is less taxing than manually creating paradigmatic property based clusters like WordNet synsets.

Barring one instance, both cluster based features outperform word based features. The reason for the drop in the accuracy of approach based on sense features for *En-PD* dataset is the domain specific nature of sentiment analysis (Blitzer et al., 2007), which is explained in the next point.

2. Domain issues are resolved while using cluster based features: For *En-PD*, the classifier developed using sense features based on paradigmatic analysis performs inferior to word based features. Compared to other datasets used for analysis, this dataset was sense annotated using an automatic WSD engine. This engine was trained on a travel domain corpus and as WSD is also domain specific, the final classification performance suffered. Additionally, as the target domain was on products, the automatic WSD engine employed had an in-domain accuracy of 78%. The sense disambiguation accuracy of the same would have lowered in a cross-domain setting. This might have had a degrading effect on the SA accuracy.

However, it was seen that classifier developed on cluster features based on syntagmatic analysis do not suffer from this. Such clusters obliterate domain relates issues. In addition, as more unlabelled data is included for clustering, the classification accuracy improves.[8] Thus, clustering may be employed to tackle other specific domain related issues in SA.

[8]It was observed that adding 0.1 million unlabelled documents, SA accuracy improved by 1%. This was observed in the case of English for which there is abundant unlabelled corpus.

3. Cluster based features using syntagmatic analysis requires lesser training data: Cluster based features drastically reduces the dimension of the feature vector. For instance, the size of sense based features for En-TD dataset was $1/6^{th}$ of the size of word based features. This reduces the perplexity of the classification model. The reduction in the perplexity leads to the reduction of training documents to attain the same classification accuracy without any dimensionality reduction. This is evident from Figure 1 where accuracy of the cluster features based on unlabelled corpora are higher even with lesser training data.

4. Effect of cluster size: The cluster size (number of clusters employed) has an implication on the purity of each cluster with respect to the application. The system performance improved upon increasing the cluster size and converged after attaining a certain level of accuracy. In general, it was found that the best classification accuracy was obtained for a cluster size between 1000 and 2500. As evident from Table 2, once the optimal accuracy is obtained, no significant changes were observed by increasing the cluster size.

5. Clustering based CLSA is effective: For target language as *Hindi*, CLSA accuracy based on cross-lingual clustering (syntagmatic) outperforms the one based on MT (refer to Table 3). This was true for the constraint clustering approach based on cross-lingual clustering. Whereas, sentiment classifier using sense (PS) or direct cluster linking (DCL) is not very effective. In case of PS approach, the coverage of the multidict was a problem. The number of a linkages between sense from *English* to *Hindi* is only around $1/3^{rd}$ the size of Princeton WordNet (Fellbaum, 1998). Similarly in case of DCL approach, monolingual likelihood is different from the cross-lingual likelihood in terms of the linkages.

6. A note on CLSA for truly resource scarce languages: Note that there is no publicly available MT system for *English* to *Marathi*. Moreover, the digital content in *Marathi* language does not have a standard encoding format. This impedes the automatic crawling of the web for corpora creation for SA. Much manual effort has to be put to collect enough corpora for analysis. However, even in these languages, unlabelled corpora is easy to obtain. Marathi was chosen to depict a truly resource scarce SA scenario. Cluster features based classifier comparatively performed well with 60% classification accuracy. An MT based system would have suffered in this case as *Marathi*, as stated earlier, is a morphologically rich language and as compared to English, has a different word ordering. This could degrade the accuracy of the machine translation itself, limiting the performance of an MT based CLSA system. All this is obliterated by the use of a cluster based CLSA approach. Moreover, as more monolingual copora is added for clustering, the cross lingual cluster linkages could be refined. This can further boost the CLSA accuracy.

8 Conclusion and Future Work

This paper explored feasibility of using word cluster based features in lieu of features based on WordNet senses for sentiment analysis to alleviate the problem of data sparsity. Abstractly, the motivation was to see if highly effective features based on paradigmatic property based clustering could be replaced with the inexpensive ones based on syntagmatic property for SA.

The study was performed for both monolingual SA and cross-lingual SA. It was found that cluster features based on syntagmatic analysis are better than the WordNet sense features based on paradigmatic analysis for SA. Invesitgation revealed that a considerable decrease in the training data could be achieved while using such class based features. Moreover, as syntagma based word clusters are homogenous, it was able to address domain specific nature of SA as well.

For CLSA, clusters linked together using unlabelled parallel corpora do away with the need of translating labelled corpora from one language to another using an intermediary MT system or bilingual dictionary. Such a method outperforms an MT based CLSA approach. Further, this approach was found to be useful in cases where there are no MT systems to perform CLSA and the language of analysis is truly resource scarce. Thus, wider implication of this study is that many widely spoken yet resource scare languages like *Pashto, Sundanese, Hausa, Gujarati* and *Punjabi* which do not have an MT system could now be analysed for sentiment. The approach presented here for CLSA will still require a parallel corpora. However, the size of the parallel corpora required

for CLSA can considerably be much lesser than the size of the parallel corpora required to train an MT system.

A naive cluster linkage algorithm based on word alignments was used to perform CLSA. As a result, there were many erroneous linkages which lowered the final SA accuracy. Better cluster-linking approaches could be explored to alleviate this problem. There are many applications which use WordNet like IR, IE *etc*. It would be interesting to see if these could be replaced by clusters based on the syntagmatic property.

References

A. R. Balamurali, Aditya Joshi, and Pushpak Bhattacharyya. 2011. Harnessing wordnet senses for supervised sentiment classification. In *Proceedings of EMNLP 2011*, pages 1081–1091, Stroudsburg, PA, USA.

A. R. Balamurali, Aditya Joshi, and Pushpak Bhattacharyya. 2012. Cross-lingual sentiment analysis for Indian languages using linked wordnets. In *Proceedings of COLING 2012*, pages 73–82, Mumbai, India.

A. R. Balamurali, Mitesh M. Khapra, and Pushpak Bhattacharyya. 2013. Lost in translation: viability of machine translation for cross language sentiment analysis. In *Proceedings of CICLing 2013*, pages 38–49, Berlin, Heidelberg.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of EMNLP 2008*, pages 127–135, Honolulu, Hawaii.

Farah Benamara, Sabatier Irit, Carmine Cesarano, Napoli Federico, and Diego Reforgiato. 2007. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *Proceedings of the International Conference on Weblogs and Social Media*.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL 2007*, pages 440–447, Prague, Czech Republic.

Julian Brooke, Milan Tofiloski, and Maite Taboada. 2009. Cross-linguistic sentiment analysis: From english to spanish. In *Proceedings of the International Conference RANLP-2009*, pages 50–54, Borovets, Bulgaria.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, pages 467–479, December.

D. Chandler. 2012. Semiotics for beginners. `http://users.aber.ac.uk/dgc/Documents/S4B/sem01.html`. Online, accessed 20-February-2013.

D. A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press.

Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of WWW 2003*, pages 519–528, New York, NY, USA.

Kevin Duh, Akinori Fujino, and Masaaki Nagata. 2011. Is machine translation ripe for cross-lingual sentiment classification? In *Proceedings of ACL-HLT 2011*, pages 429–433, Stroudsburg, PA, USA.

Manaal Faruqui and Sebastian Padó. 2010. Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Hatem Ghorbel and David Jacot. 2011. Further experiments in sentiment analysis of french movie reviews. In *Proceedings of AWIC 2011*, pages 19–28, Fribourg, Switzerland.

Gholamreza Haffari, Marzieh Razavi, and Anoop Sarkar. 2011. An ensemble model that combines syntactic and semantic clustering for discriminative dependency parsing. In *Proceedings of ACL-HLT 2011*, pages 710–714, Stroudsburg, PA, USA.

Kanayama Hiroshi, Nasukawa Tetsuya, and Watanabe Hideo. 2004. Deeper sentiment analysis using machine translation technology. In *Proceedings of COLING 2004*, Stroudsburg, PA, USA.

Daisuke Ikeda, Hiroya Takamura, Lev arie Ratinov, and Manabu Okumura. 2008. Learning to shift the polarity of words for sentiment classification. In *Proceedings of the Third International Joint Conference on Natural Language Processing*.

Mitesh Khapra, Sapan Shah, Piyush Kedia, and Pushpak Bhattacharyya. 2010. Domain-specific word sense disambiguation combining corpus based and wordnet based parameters. In *Proceedings of Global Wordnet Conference*.

Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of ACL-HLT 2008*, pages 595–603, Columbus, Ohio.

Percy Liang. 2005. Semi-supervised learning for natural language. M. eng. thesis, Massachusetts Institute of Technology.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of CIKM 2009*, pages 375–384, New York, NY, USA.

Bin Lu, Chenhao Tan, Claire Cardie, and Benjamin K. Tsou. 2011. Joint bilingual sentiment classification with unlabeled parallel corpora. In *Proceedings of ACL-HLT 2011*, pages 320–330, Stroudsburg, PA, USA.

Sven Martin, Jrg Liermann, and Hermann Ney. 1995. Algorithms for bigram and trigram word clustering. In *Speech Communication*, pages 1253–1256.

Justin Martineau and Tim Finin. 2009. Delta TFIDF: An improved feature space for sentiment analysis. In *Proceedings of ICWSM*.

Shotaro Matsumoto, Hiroya Takamura, and Manabu Okumura. 2005. Sentiment classification using word sub-sequences and dependency sub-trees. In *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, pages 301–311.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of WWW 2007*, pages 171–180, New York, NY, USA.

Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In *Proceedings of HLT-NAACL 2004: Main Proceedings*, pages 337–342, Boston, Massachusetts, USA.

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. 2007. Generating complex morphology for machine translation. In *Proceedings of ACL 2007*, pages 128–135, Prague, Czech Republic.

Rajat Mohanty, Pushpak Bhattacharyya, Prabhakar Pande, Shraddha Kalele, Mitesh Khapra, and Aditya Sharma. 2008. Synset based multilingual dictionary: Insights, applications and challenges. In *Proceedings of Global Wordnet Conference*.

Tony Mullen and Nigel Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of EMNLP 2004*, pages 412–418, Barcelona, Spain.

Tetsuji Nakagawa, Kentaro Inui, and Sadao Kurohashi. 2010. Dependency tree-based sentiment classification using crfs with hidden variables. In *Proceedings of HLT-NAACL 2010*, pages 786–794, Stroudsburg, PA, USA.

Vincent Ng, Sajib Dasgupta, and S. M. Niaz Arifin. 2006. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *Proceedings of the COLING 2006*, pages 611–618, Stroudsburg, PA, USA.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

Bo Pang and Lillian Lee. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP 2002*, pages 79–86, Stroudsburg, PA, USA.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, January.

Vassiliki Rentoumi, George Giannakopoulos, Vangelis Karkaletsis, and George A. Vouros. 2009. Sentiment analysis of figurative language using a word sense disambiguation approach. In *Proceedings of RANLP 2009*, pages 370–375, Borovets, Bulgaria, September.

Niall Rooney, Hui Wang, Fiona Browne, Fergal Monaghan, Jann Mller, Alan Sergeant, Zhiwei Lin, Philip Taylor, and Vladimir Dobrynin. 2011. An exploration into the use of contextual document clustering for cluster sentiment analysis. In *Proceedings of RANLP 2011*, pages 140–145, Hissar, Bulgaria.

C. Samuelsson and W. Reichl. 1999. A class-based language model for large-vocabulary speech recognition extracted from part-of-speech statistics. In *Proceedings of ICASSP 1999*, pages 537–540.

Sara Stymne. 2012. Clustered word classes for preordering in statistical machine translation. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, pages 28–34.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual Word Clusters for Direct Transfer of Linguistic Structure. In *Proceedings of NAACL-HLT 2012*, pages 477–487, Montréal, Canada.

Martin Tamara, Balahur Alexandra, and Montoyo Andres. 2010. Word sense disambiguation in opinion mining: Pros and cons. *Journal Research in Computing Science*, 46:119–130.

Stephen Tratz and Eduard Hovy. 2011. A fast, accurate, non-projective, semantically-enriched parser. In *Proceedings of EMNLP 2011*, pages 1257–1268, Stroudsburg, PA, USA.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of ACL 2010*, pages 384–394, Stroudsburg, PA, USA.

Peter D. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL 2002*, pages 417–424, Stroudsburg, PA, USA.

Jakob Uszkoreit and Thorsten Brants. 2008. Distributed word clustering for large scale class-based language modeling in machine translation. In *Proceedings of ACL-HLT 2008*, pages 755–762, Columbus, Ohio.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of ACL 2009*, pages 235–243, Stroudsburg, PA, USA.

Gui-Rong Xue, Chenxi Lin, Qiang Yang, WenSi Xi, Hua-Jun Zeng, Yong Yu, and Zheng Chen. 2005. Scalable collaborative filtering using cluster-based smoothing. In *Proceedings of SIGIR 2005*, pages 114–121, New York, NY, USA.

Qiang Ye, Ziqiong Zhang, and Rob Law. 2009. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, 36(3, Part 2):6527–6535.

Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. 2011. Clustering product features for opinion mining. In *Proceedings of WSDM 2011*, pages 347–354, New York, NY, USA.

Yue Zhang and Joakim Nivre. 2011. Transition-based dependency parsing with rich non-local features. In *Proceedings of ACL-HLT 2011*, pages 188–193, Stroudsburg, PA, USA.