Cross-Lingual Sentiment Analysis for Indian Languages using Linked WordNets

Balamurali A R[1,2] *Aditya Joshi*[1] *Pushpak Bhattacharyya*[1]
(1) Indian Institute of Technology, Mumbai, India 400076
(2) IITB-Monash Research Academy, Mumbai, India 400076
balamurali@iitb.ac.in, aditya.jo@iitb.ac.in, pb@iitb.ac.in

ABSTRACT
Cross-Lingual Sentiment Analysis (CLSA) is the task of predicting the polarity of the opinion expressed in a text in a language L_{test} using a classifier trained on the corpus of another language L_{train}. Popular approaches use Machine Translation (MT) to convert the test document in L_{test} to L_{train} and use the classifier of L_{train}. However, MT systems do not exist for most pairs of languages and even if they do, their translation accuracy is low. So we present an alternative approach to CLSA using WordNet senses as features for supervised sentiment classification. A document in L_{test} is tested for polarity through a classifier trained on *sense marked* and polarity labeled corpora of L_{train}. The crux of the idea is to use the linked WordNets of two languages to bridge the language gap. We report our results on two widely spoken Indian languages, Hindi (450 million speakers) and Marathi (72 million speakers), which do not have an MT system between them. The sense-based approach gives a CLSA accuracy of 72% and 84% for Hindi and Marathi sentiment classification respectively. This is an improvement of 14%-15% over an approach that uses a bilingual dictionary.

KEYWORDS: Sentiment Analysis, Cross Lingual Sentiment Analysis, Linked Wordnets, Semantic Features, Sense Space.

1 Introduction

Sentiment Analysis (SA) is the task of inferring polarity of an opinion in a text. Though the majority of the work in SA is for English, there has been work in other languages as well such as Chinese, Japanese, German and Spanish (Seki et al., 2007; Nakagawa et al., 2010; Schulz et al., 2010). To perform SA on these languages, cross-lingual approaches are often used due to the lack of annotated content in these languages. In Cross-Lingual Sentiment Analysis (CLSA), the training corpus in one language (call it L_{train}) is used to predict the sentiment of documents in another language (call it L_{test}). Machine Translation is often employed for CLSA (Wan, 2009; Wei and Pal, 2010). A document in L_{test} is translated into L_{train} and is checked for polarity using the classifier trained on the polarity marked documents of L_{train}. However, MT is resource-intensive and does not exist for most pairs of languages.

WordNet (Fellbaum, 1998) is a widely used lexical resource in the NLP community and is present in many languages.[1] Most of the WordNets are developed using the expansion based approach (Vossen, 1998; Bhattacharyya, 2010) wherein a new WordNet for a target language (L_t) is created by adding words which represent the corresponding synsets in the source language (L_s) WordNet. As a consequence, corresponding concepts in L_s and L_t have the same synset (concept) identifier. Our work leverages this fact, and uses WordNet senses as features for building a classifier in L_{train}. The document to be tested for polarity is preprocessed by replacing words in this document with the corresponding synset identifiers. This step eliminates the distinction between L_{train} and L_{test} as far as the document is concerned. The document vector created from the sense-based features could belong to any language. The preprocessed document is then given to the classifier coming from L_{train} for polarity detection.

This work is an extension our sense-based SA work on English (Balamurali et al., 2011) where we showed that *WordNet synset-based features perform better than word-based features for sentiment analysis*. Here, we carry out our study on two widely spoken Indian languages: *Hindi* and *Marathi*. These languages belong to the Indo-Aryan subgroup of the Indo-European language family. For these two languages, we first verify the superiority of sense-based features over word-based features for SA. Thereafter we proceed to verify the efficacy of the sense-based approach for cross-lingual sentiment analysis for these two languages. This work differs from existing works(Brooke et al., 2009; Wan, 2009; Wei and Pal, 2010; Banea et al., 2008) on CLSA in two aspects: (i) our focus is not necessarily to use a resource-rich language to help a resource-scarce language but *can be applied to any two languages which share a common sense space* (by using WordNets with matching synset identifiers); (ii) our work is an alternative to *MT-based cross-lingual sentiment analysis* for languages which do not have an MT system between them.

2 Background Study: Word Senses for SA

In our previous work (Balamurali et al., 2011), we showed that word senses act as better features than lexeme-based features for document level SA. We termed this feature space as *synset space* or *sense space*. In the sense space, the semantics of document is represented in a compact way using synset identifiers.

Different variants of a travel review domain corpus are generated by using automatic/manual sense disambiguation techniques. Thereafter, classification accuracy of classifiers based on

[1] http://www.globalWordNet.org/gwa/WordNet_table.html

different sense-based and word-based features were compared. The experimental results show that *WordNet senses act as better features compared to words alone.*

The following subsection validates this hypothesis for Hindi and Marathi. Since the documents for training and testing belong to the same language, we refer to this set of classification experiments as *in-language sentiment classification*.

2.1 WordNet Senses as Better Features: Approach

A classifier is trained for each of the following feature representations: Words (*W*), Manually annotated word senses (*M*), Automatically annotated word senses (*I*), Words and *manually* annotated word senses (*W+S(M)*) and Words and *automatically* annotated word senses (*W+S(I)*). At present, the development of Hindi and Marathi WordNets is not complete. Thus, a number of words belonging to open POS categories (*e.g.* nouns) do not have corresponding synsets created. We used W+S(M) and W+S(I) representations in order to alleviate problems that can arise due to these missing synsets.

We perform our experiments on the above feature representations for *in-language sentiment classification* and compare their performance. The results are discussed in section 6.1.

3 Word Senses for Cross-Lingual SA

We now describe our approach to cross-lingual SA, which is the focus of this work. This approach harnesses word senses to build a supervised sentiment classifier in a cross-lingual setting (*i.e.,* when the L_{train} and L_{test} are different).

Our baseline as well as sense-based approach center around the WordNets of the two languages *viz.*, Hindi and Marathi. WordNets of Hindi and Marathi have been developed using an expansion approach. This approach involves expanding the Marathi WordNet by adding concept definition for concepts from Hindi WordNet. Subsequently, corresponding related terms are added and mapped. Thus, corresponding concepts/synsets in WordNets of both languages have the same synset identifier. Once this mapping is completed, concepts found only in the target language are added.

An instance of WordNets which are collectively developed for multiple languages is referred to as Multidict (Mohanty et al., 2008). In a Multidict, each row constitutes a concept, identified by a synset identifier.

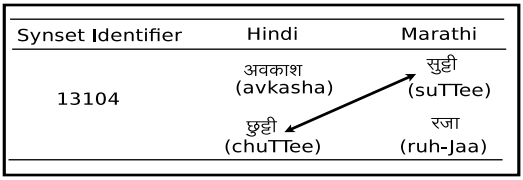

Figure 1: An example entry *(concept: holiday)* in Multidict for Hindi and Marathi

Each column contains synonymous terms representing these concepts in different languages. Further, a manual cross link is provided between words in one language to another based on their lexical preference.

The words in the corresponding synsets are thus translations of each other in specific contexts. For example, an entry pertaining to Marathi and Hindi can be explained as follows (Figure 1):

13104 (Synset identifier) pertains to the concept of *holiday* and its related terms are *suTTee* and *ruh-jaa* in Marathi and *chuTTee* and *avkasha* in Hindi. The cross links shown in the above entry indicates that when the Marathi word *suTTee* is used in the sense represented by the synset identifier *13104*, its exact Hindi translation is *chuTTee* (*i.e.,* this translation is more preferred over the other related Hindi words of the same synset).

3.1 Our Approach: Sense-based Representation

Following the fact that the Hindi and Marathi WordNet have the same synset identifier for the same concept, we represent words in the two languages by corresponding synset identifiers.

Thus, in a cross-lingual setting for a given target language, we map the words of the training as well as the test corpus to their WordNet synset identifiers. A classification model is learnt on the training corpus and tested on the test corpus. Both corpora consists of synset identifiers. This experiment is performed for two variants of the corpora: one with manually annotated senses and another with automatically annotated senses. Thus, in the context of using senses as features for cross-lingual sentiment analysis, we evaluate the following approaches: 1. A group of word senses that have been manually annotated (M), 2. A group of word senses that have been annotated by an automatic Word Sense Disambiguation (WSD) engine (I).

The replacement of a word by its synset identifier is carried out for all documents in the training corpus and the test corpus. The representation of the new corpora is in a common feature space, *i.e.,* the sense space.

3.2 Baseline: Naïve Translation Using Lexeme Replacement

MT-based techniques have been the main way of performing cross-lingual SA (Wan, 2009; Wei and Pal, 2010). The obvious choice for a baseline to compare our approach would have been a MT based CLSA approach. However, at present, there exists no Hindi-Marathi MT system. Hence we develop a strategy for obtaining a naïve translation of the corpus-based on lexical transfer which forms the baseline for comparing sentiment classification accuracy of the proposed cross-lingual SA based on synset representation.

Our approach consists of converting a document from the L_{test} to the L_{train} so that a classifier modeled on documents from the training language can be used. The words in the test documents are mapped to the corresponding words in the training language to obtain a naïve translation. No semantic/syntactic transfer is maintained. We use Multidict to translate synonymous terms in different languages, namely Hindi and Marathi (Mohanty et al., 2008). We offer two versions which differ from each other based on the *replacement lexeme chosen*.

Exact word replacement (E): Based on the disambiguated sense identifier, the exact cross-linked word from the source language is used for the replacement. Hence, for the word *suTTee*, the translation *chuTTee* will be selected (Figure 1).

Random word replacement (R): Based on the disambiguated sense identifier, the cross linked word from the source language is used for the replacement. This word in Figure 1 is not necessarily the exact (*preferred*) translation as mentioned above. For example, for the word *suTTee*, some random translation from the same synset will be selected, for example *ruh-jaa*, instead of the preferred translation *chuTTee* (Figure 1) will be selected.

The replacement is carried out for all documents in the test corpus (originally in L_{test}) to generate a new test corpus (containing words in L_{train}). We understand this naïve translation

may not give as strong a baseline as a statistical MT-based approach, but given the state of these languages, we believe the results obtained are fairly comparable.

4 Datasets

The dataset we created for Hindi and Marathi consists of user-written travel destination reviews. We collected them from various blogs and Sunday travel editorials. A review consists of approximately 4-5 sentences of 10-15 words each. The Hindi corpus consists of approximately 100 positive and 100 negative reviews while the Marathi corpus consists of approximately 75 positive and 75 negative reviews. The documents are labeled with polarity (positive/negative) by a native speaker.

To create the manual sense-annotated corpus, the words were manually annotated by a native speaker. Based on the word and POS category, the annotation tool shows all possible sense entries for that word in the WordNet. The lexicographer then chooses the right sense based on the context. Hindi corpora contains 11038 words whereas Marathi corpora contains 12566 words. To generate automatic sense-annotated corpus, we use the engine based on the IWSD algorithm, which is trained on the tourism domain and can operate on Hindi, Marathi and English. We chose the travel review domain for our analysis because the IWSD engine was trained on this domain.

POS	#Words	Precision	Recall	F-score
Noun	2601	73.26%	70.59%	71.90%
Adverb	506	80.08%	79.45%	79.76%
Adjective	700	56.65%	54.14%	55.37%
Verb	1487	54.11%	51.78%	52.92%
Overall	5294	66.41%	63.98%	65.17%

Table 1: Annotation statistics for Hindi

POS	#Words	Precision	Recall	F-score
Noun	1628	76.60%	75.80%	76.20%
Adverb	204	73.53%	73.53%	73.53%
Adjective	583	76.27%	74.96%	75.61%
Verb	363	82.35%	80.99%	81.67%
Overall	2778	77.05%	76.13%	76.59%

Table 2: Annotation statistics for Marathi

Tables 1 and 2 show the evaluation of sense disambiguation statistics for IWSD for Hindi and Marathi respectively.

5 Experimental Setup

The experiments are performed using C-SVM (linear kernel with default parameters; C=0.0, ϵ=0.0010) available as a part of LibSVM package.[2] We chose SVM as its known to be a good learner for sentiment classification (Pang and Lee, 2002).

To conduct experiments on words as features, we perform stop-word removal and word stemming. For synset-based experiments, words in the corpus are substituted with synset identifiers along with POS categories, which are used as features. To create automatically sense-annotated corpora, we use the state-of-the-art domain specific word sense disambiguation (IWSD) algorithm by Khapra et al. (2010) for sense disambiguating our datasets in the two languages.

The results are evaluated using commonly used classification metrics: classification accuracy, Fscore, recall and precision. Recall and precision for each polarity label is also calculated for analysis.

For our background study experiments pertaining to the in-language sentiment classification, a two-fold validation of five repeats is carried out. Each repeat consists of a random configuration

[2] http://www.csie.ntu.edu.tw/c̃jlin/libsvm

of test/train documents maintained across different representations for a given run. Such a cross-fold validation is taken to minimize the variance between the classification results of different folds since the sizes of the corpus are not that large (Dietterich, 1998).

6 Results and Discussions

Our results are divided into two parts. Section 6.1 shows the results related to our background study pertaining to in-language sentiment classification. In section 6.2, we compare the approaches for cross-lingual sentiment analysis.

6.1 In-language Classification

The results of in-language classification for Marathi and Hindi are shown in Table 3[3]. We consider unigram words as the baseline (Words) for comparison. Note that since cross-lingual SA using 'perfect' translation from target to source language is identical to in-language sentiment classification, these results act as an *upper bound/skyline* to the performance of cross-lingual SA. While using sense-based features, we also use the POS information and hence to have a fair comparison, we use an additional baseline which include the POS information in addition to unigram features (represented as *Words + POS)*.

L_{train} & L_{test} : **Marathi**							
Feature Representation	**Accuracy**	**PF**	**NF**	**PP**	**NP**	**PR**	**NR**
Words(Baseline)	86.53	85.13	86.96	96.68	80.25	76.05	94.90
Words + POS (Baseline)	83.32	79.91	85.42	97.00	76.92	69.33	97.00
Sense (M)	97.45	97.38	97.62	100.00	95.36	94.89	100.00
Sense + Words (M)	**97.87**	97.82	97.94	100.00	95.97	95.74	100.00
Sense(I)	93.44	93.97	92.94	89.25	99.19	99.21	87.43
Sense + Words (I)	92.78	93.35	92.32	88.14	99.17	99.20	86.36
L_{train} & L_{test} : **Hindi**							
Feature Representation	**Accuracy**	**PF**	**NF**	**PP**	**NP**	**PR**	**NR**
Words(Baseline)	65.64	61.65	64.83	71.38	62.29	54.25	67.60
Words+POS(Baseline)	76.34	70.18	79.92	89.42	70.34	58.27	92.80
Sense(M)	82.57	78.55	84.45	89.68	78.34	69.88	91.60
Words+Sense(M)	**83.06**	79.48	85.09	92.11	77.86	69.90	93.80
Sense(I)	81.92	78.00	83.25	88.63	78.98	69.65	88.00
Words+Sense(I)	81.21	78.03	83.50	89.35	77.29	69.26	90.80

Table 3: Background study: In-language sentiment classification showing the skyline performance for Marathi and Hindi; PF-Positive F-score, NF-Negative F-score, PP-Positive Precision(%), NP-Negative Precision(%), PR-Positive Recall (%), NR-Negative Recall (%)

Overall Sentiment Classification:

All sense-based features give a higher overall accuracy than the baseline for both Marathi and Hindi. The baseline for Hindi is lower than that for Marathi. However, manually annotated sense-based features perform better than the baseline by 11.3% for Marathi and 6.7% for Hindi. The classification accuracy of the combination of manually annotated synsets and words is comparable to that of manually annotated synsets for both the languages.

As expected, automatic sense disambiguation-based features perform better than the baseline but lower than manually annotated features. For Marathi, the classification accuracy for

[3]All results statistically significant (paired-T test, confidence=95%) with respect to the baseline. 3. For Marathi, Sense (M) and Words + Sense (M) results are not significant. Same is the case for Sense (I) and Words + Sense (I) for Hindi.

automatic sense disambiguation-based representation degrades by 4% below the manually annotated counterpart. This degradation is less significant in case of Hindi as the overall accuracy of Hindi sense disambiguation engine is only 66% (refer to Table 1). This suggests that even a low accuracy sense disambiguation may be sufficient to obtain better results than word based features.

6.2 CLSA Accuracy

L_{train} : **Hindi** & L_{test} : **Marathi**							
Feature Representation	**Accuracy**	**PF**	**NF**	**PP**	**NP**	**PR**	**NR**
Words(E) Baseline 1	71.64	72.22	62.86	75.36	67.69	69.33	58.67
Words(R) Baseline 2	70.15	71.23	60.87	73.24	66.67	69.33	56.00
Senses(M)	84.00	81.54	85.88	96.36	76.84	70.67	97.33
Senses(I)	**84.50**	83.33	85.51	96.15	76.62	73.53	96.72
L_{train} : **Marathi** & L_{test} : **Hindi**							
Feature Representation	**Accuracy**	**PF**	**NF**	**PP**	**NP**	**PR**	**NR**
Words(E) Baseline 1	56.42	29.31	64.37	94.44	52.17	17.35	84.00
Words(R) Baseline 2	57.69	30.77	66.16	94.74	53.37	18.37	87.00
Senses(M)	**72.08**	62.82	77.18	87.50	65.96	49.00	93.00
Senses(I)	68.11	61.04	72.81	77.05	63.71	50.54	84.95

Table 4: Cross-Lingual sentiment classification for target languages Marathi and Hindi; PF-Positive F-score, NF-Negative F-score, PP-Positive Precision(%), NP-Negative Precision(%), PR-Positive Recall (%), NR-Negative Recall (%)

Sense based CLSA accuracy along with the baseline accuracy is shown in Table 4[4].

L_{test} **- Marathi:** In-language classification accuracy for Marathi using words as features is only 86.53% (refer to Table 3). In a way, this forms the upper bound for a perfectly translated document. In the case of the naïve translation-based approach, an accuracy of 71.64% and 70.15% for Words (E) and Words (R) is obtained respectively. Both the manually and the automatically annotated sense-based features show an improvement of 12% (approximately) over both the baselines.

L_{test} **- Hindi:** When Hindi is the target language, the baseline using lexeme replacement is lower than the baseline for Marathi. An approximate 15% improvement over the baseline is observed for manually annotated sense-based features (which has an accuracy of 72%). Sense-based features developed using automatic sense disambiguation work with a lower accuracy with respect to manually annotated synsets.

A considerable improvement in the positive recall can be seen for Hindi as the target language. The same can be said about the negative precision. These results highlight the effectiveness of synsets as features for negative sentiment detection in a cross-lingual setup.

As most of the Indian languages do not have MT systems between them, we believe this approach can be an alternative to MT based CLSA approaches. Our approach is at par with MT based CLSA approach as our results are not far behind the in-language classification results. Hence MT based CLSA approaches are comparable with our approach as they too fall behind in-language classification results (based on the results of an independent study).

[4] All results are statistically significant with respect to the baseline. However, baseline 1 and baseline 2 are not statistically significant and so is the case for Sense (M) and Sense(I) accuracy figures for Marathi (as L_{test})

Effect of Automatic WSD on Classification Accuracy

Sense annotation accuracy (Fscore) of the WSD engine used for annotating the words with their respective sense is 65% and 76% (Tables 1 and 2) for Hindi and Marathi respectively. Annotation accuracy is less for Hindi as there are more finer senses in Hindi WordNet than in Marathi WordNet. Thus, there is a higher chance of assigning an incorrect sense for a word in Hindi than compared to a word in Marathi. However, the fall in classification accuracy due to this reason is not reflected on the in-language sentiment classification accuracy of Hindi and Marathi respectively. Nevertheless, there is a drop in the cross lingual accuracy when L_{test} is Hindi, which may be due to relatively small training corpora size of Marathi when compared to Hindi. Marathi corpus is half the size of Hindi corpus and hence contain less training samples where L_{test} is Hindi. As both the manually and the automatically assigned sense based features give almost similar cross lingual accuracy for the case when L_{test} is Hindi, we strongly believe that classification accuracy can be improved by adding more Marathi documents.

7 Error Analysis

Two possible reasons for errors in the existing approach that we found are:

1. Missing Concepts: As the Marathi WordNet is created using the expansion approach from the Hindi WordNet, almost all concepts present in the Marathi WordNet are derived from the Hindi WordNet. In contrast, there are many concepts present in the Hindi WordNet but not yet included in the Marathi WordNet. This leads to a low cross-lingual sentiment classification accuracy using sense-based features with target language as Hindi.

2. Hindi Morph Analyzer Defect: The accuracy of sense-based in-language classification for Hindi is comparatively lower than that for Marathi. We traced the problem to the sense annotation tool used by the manual annotator. The morphological analyzer used to find the root word (for verbs) did not match Hindi WordNet entries for verb synsets in many cases, thus reducing the coverage of the annotation.

8 Conclusion and Future Work

We presented an approach to cross-lingual SA that uses WordNet synset identifiers as features of a supervised classifier. Our sense-based approach provides a cross-lingual classification accuracy of 72% and 84% for Hindi and Marathi respectively, which is an improvement of 14% - 15% over the baseline based on a cross-lingual approach using a naïve translation of the training and test corpus. We also performed experiments based on a sense marked corpora using an automatic WSD engine. Results suggest that even a low quality word sense disambiguation leads to an improvement in the performance of sentiment classification. In summary, we have shown that WordNet synsets can act as good features for cross-lingual SA.

In future, we would like to perform sentiment analysis in a multilingual setup. Training data belonging to multiple languages can be leveraged to perform SA for some specific target language. Additionally, we would like to compare our CLSA approach with a MT based approach. For this, we plan to perform same set of experiments for languages (like English and Romanian) which have a linked wordnet as well a MT system between them.

References

Balamurali, A., Joshi, A., and Bhattacharyya, P. (2011). Harnessing wordnet senses for supervised sentiment classification. In *Proc. of EMNLP-11*, pages 1081–1091.

Banea, C., Mihalcea, R., Wiebe, J., and Hassan, S. (2008). Multilingual subjectivity analysis using machine translation. In *Proc. of EMNLP-08*, pages 127–135.

Bhattacharyya, P. (2010). Indowordnet. In *Proc. of LREC-10*, Valletta, Malta. European Language Resources Association (ELRA).

Brooke, J., Tofiloski, M., and Taboada, M. (2009). Cross-Linguistic Sentiment Analysis: From English to Spanish. In *Proc. of RANLP-09*.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.

Khapra, M., Shah, S., Kedia, P., and Bhattacharyya, P. (2010). Domain-specific word sense disambiguation combining corpus basedand wordnet based parameters. In *Proceedings of GWC-10*.

Mohanty, R., Bhattacharyya, P., Pande, P., Kalele, S., Khapra, M., and Sharma, A. (2008). Synset based multilingual dictionary: Insights, applications and challenges. In *Proc. of GWC-08*.

Nakagawa, T., Inui, K., and Kurohashi., S. (2010). Dependency tree-based sentiment classification using crfs with hidden variables. In *Proc. of NAACL/HLT-10*.

Pang, B. and Lee, L. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proc. of EMNLP-02*, pages 79–86.

Schulz, J. M., Womser-Hacke, C., and Mandl, T. (2010). Multilingual corpus development for opinion mining. In *Proc. of LREC-10*.

Seki, Y., Evans, D. K., Ku, L.-W., Sun, L., Chen, H.-H., and Kando, N. (2007). Overview of multilingual opinion analysis task at ntcir-7. In *Proc. of NTCIR-7 Workshop*.

Vossen, P. (1998). Eurowordnet: a multilingual database with lexical semantic networks. In *International Conference on Computational Linguistics*.

Wan, X. (2009). Co-training for cross-lingual sentiment classification. In *Proc. of ACL-AFNLP-09*, pages 235–243.

Wei, B. and Pal, C. (2010). Cross lingual adaptation: an experiment on sentiment classifications. In *Proc. of ACL-10*, pages 258–262.