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 Bombay (2) IITB-Monash Research Academy

Introduction

Cross-Lingual Sentiment Analysis (CLSA)

Predicting sentiment polarity of text in language L_{test} using a classifier trained on corpus of language L_{train}.

Existing CLSA Approaches:

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.

Objective:

To perform CLSA for languages which do not have MT systems between them but have linked WordNets.

Background

Word Senses for SA:

Word senses ***act as better*** features than lexeme-based features for document level SA.

We term this feature space as *synset space or sense space.*

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)).*

Testing Hypothesis: Compare the accuracies of various sense based classifiers with word based classifier

Language of Analysis

Hindi and Marathi

Hindi and Marathi lbelongs to the Indo-Aryan subgroup of the Indo-European language family. Marathi Wordnet has been developed from Hindi Wordnet 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 mapped.

Synset Identifier	Hindi	Marathi
13104	अवकाश (avkasha)	सुट्टी (suTTee)
	छुट्टी (chuTTee)	रजा (ruh-Jaa)

An instance of WordNets collectively developed for multiple languages is referred to as Multidict. In a Multidict, each row constitutes a concept, identified by a synset identifier. Each column contains synonymous terms representing these concepts in different language. Further, a manual cross link is provided between words in one language to another based on their lexical preference.

Approach

Map words in training and corpus to corresponding Wordnet synset identifiers. A classification model is learnt using synset identifiers as features.

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. Though train and test languages are different, their representation for the classifier is in a common feature space, i.e., the sense space.

Baseline:

- No MT system for Marathi-Hindi exists.
- A Naïve translation based on lexical transfer is used as baseline.
  - Use Multidict to translate synonymous terms in different languages

- **Exact word replacement (E)**: Based on the disambiguated sense identifier, the exact cross-linked word from the source language is produced as translated word.

- **Random word replacement (R):** Based on the disambiguated sense identifier, a random word from cross-linked synset in Wordnet of source language is used for replacement.

Datasets and Experimental Setup

The Hindi travel review corpus (11038) consists of100 positive and 100 negative reviews while the Marathi travel review corpus (12566) consists of approximately 75 positive and 75 negative reviews.

Sense annotation:

To create manual sense-annotated corpus, words were manually annotated by a native speaker. To generate automatic sense-annotated corpus, we use IWSD engine that has been trained on the tourism domain.

The experiments are performed using C-SVM (linear kernel with default parameters; C=0.0,=0.0010) available as a part of LibSVM package.

Results

In-Language Sentiment Analysis Results

L-train & L-test: Marathi

Feature Representation	Accuracy	Positive Fscore	Negative Fscore	Postive Precision	Negative Precision	Positive Recall	Negative Recall
Words(Baseline)	86.53	85.13	86.96	96.68	80.25	76.05	94.9
Words + POS (Baseline)	83.32	79.91	85.42	97	76.92	69.33	97
Sense (M)	97.45	97.38	97.62	100	95.36	94.89	100
Sense + Words (M)	97.87	97.82	97.94	100	95.97	95.74	100
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	99.2	86.36

L-train & L-test: Hindi

Feature Representation	Accuracy	Positive Fscore	Negative Fscore	Postive Precision	Negative Precision	Positive Recall	Negative Recall
Words(Baseline)	65.64	61.65	64.83	71.38	62.29	54.25	67.6
Words+POS(Baseline)	76.34	70.18	79.92	89.42	70.34	58.27	92.8
Sense(M)	82.57	78.55	84.45	89.68	78.34	69.88	91.6
Words+Sense(M)	83.06	79.48	85.09	92.11	77.86	69.9	93.8
Sense(I)	81.92	78	83.25	88.63	78.98	69.65	88
Words+Sense(I)	81.21	78.03	83.5	89.35	77.29	69.26	90.8

Cross-Lingual Sentiment Analysis Results

L-train: Hindi & L-test: Marathi

Feature Representation	Accuracy	Positive Fscore	Negative Fscore	Postive Precision	Negative Precision	Positive Recall	Negative Recall
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
Senses(M)	84	81.54	85.88	96.36	76.84	70.67	97.33
Senses(I)	84.5	83.33	85.51	96.15	76.62	73.53	96.72

L-train: Marathi & L-test: Hindi

Feature Representation	Accuracy	Positive Fscore	Negative Fscore	Postive Precision	Negative Precision	Positive Recall	Negative Recall
Words(E) Baseline 1	56.42	29.31	64.37	94.44	52.17	17.35	84
Words(R) Baseline 2	57.69	30.77	66.16	94.74	53.37	18.37	87
Senses(M)	72.08	62.82	77.18	87.5	65.96	49	93
Senses(I)	68.11	61.04	72.81	77.05	63.71	50.54	84.95

Error Analysis

1. **Missing Concepts**: Many concepts present in the Hindi Wordnet but not yet included in the Marathi Wordnet.

2. **Hindi Morph Analyzer Defect:** Reduced sense coverage due to verb detection problem

Future Work

- Training on data belonging to multiple languages

- Compare with MT based CLSA system