VERB BASED MANIPURI SENTIMENT ANALYSIS

Kishorjit Nongmeikapam[1], Dilipkumar Khangembam[1], Wangkheimayum Hemkumar[1], Shinghajit Khuraijam[2] and Sivaji Bandyopadhyay[3]

[1]Department of Computer Engineering, Manipur Institute of Technology, Manipur University, Imphal, India
[2]Senior Software Enggineer at Accenture, India
[3]Department of Computer Engineering, Jadav University, Kolkata, India

ABSTRACT

This paper deals about the sentiment analysis of the Manipuri article. The language is very highly agglutinative in Nature. The document files are the letters to the editor of few local daily newspapers. The text is processed for Part of Speech (POS) tagging using Conditional Random Field (CRF). The lexicon of verbs is modified with the sentiment polarity (Positive or Negative or Neutral) manually. With the POS tagger the verbs of each sentence are identified and the modified lexicon of verbs is used to notify the polarity of the sentiment in the sentence. The total number of polarity for each category that is positive, negative and neutral is counted separately. The highest total of the three is the deciding factor of the sentiment polarity of the document. The system shows a recall of 72.10%, a precision of 78.14% and a F-measure of 75.00%.

KEYWORDS

CRF, POS, Sentiment, Polarity, Manipuri

1. INTRODUCTION

This document describes, and is written to conform to, author guidelines for the journals of AIRCC series. It is prepared in Microsoft Word as a .doc document. Although other means of preparation are acceptable, final, camera-ready versions must conform to this layout. Microsoft Word terminology is used where appropriate in this document. Although formatting instructions may often appear daunting, the simplest approach is to use this template and insert headings and text into it as appropriate.

2. RELATED WORKS

The work of sentiment analysis in Manipuri is not reported up to the best of the author's knowledge. Works on emotion identification is reported in [2] and [3]. For Indian languages works on SentiWordNet and subjectivity detection is reported in [4] and [5]. CRF based work on sentiment analysis is reported in [6].

3. CONCEPT OF CONDITION RANDOM FIELD

The concept of Conditional Random Field [7] is developed in order to calculate the conditional probabilities of values on other designated input nodes of undirected graphical models. CRF encodes a conditional probability distribution with a given set of features. It is an unsupervised approach where the system learns by giving some training and can be used for testing other texts.

The conditional probability of a state sequence $X=(x_1, x_2,..x_T)$ given an observation sequence $Y=(y_1, y_2,..y_T)$ is calculated as :

$$P(Y|X) = \frac{1}{Z_X} \exp\left(\sum_{t=1}^{T} \sum_{k} \lambda_k f_k(y_{t-1}, y_t, X, t)\right) \quad ---(1)$$

where, $f_k(y_{t-1}, y_t, X, t)$ is a feature function whose weight λ_k is a learnt weight associated with $f_k$and to be learned via training. The values of the feature functions may range between $-\infty \ldots +\infty$, but typically they are binary. Z_X is the normalization factor:

$$Z_X = \sum_{y} \exp \sum_{t=1}^{T} \sum_{k} \lambda_k f_k(y_{t-1}, y_t, X, t)) \quad ---(2)$$

which is calculated in order to make the probability of all state sequences sum to 1. This is calculated as in Hidden Markov Model (HMM) and can be obtained efficiently by dynamic programming. Since CRF defines the conditional probability P(Y|X), the appropriate objective for parameter learning is to maximize the conditional likelihood of the state sequence or training data.

$$\sum_{i=1}^{N} \log P(y^i \mid x^i) \quad ---(3)$$

where, $\{(x^i, y^i)\}$ is the labeled training data.
Gaussian prior on the λ's is used to regularize the training (i.e., smoothing). If $\lambda \sim N(0,\rho^2)$, the objective function becomes,

$$\sum_{i=1}^{N} \log P(y^i \mid x^i) - \sum_{k} \frac{\lambda_i^2}{2\rho^2} \quad ---(4)$$

The objective function is concave, so the λ*'s*have a unique set of optimal values.

4. SYSTEM DESIGN

Fig. 1 gives the brief block diagram of the system. The first step of system works with identifying of verbs after running the Part of Speech (POS) tagger using CRF. This is because the sentiment of the sentence is highly dependent on the verbs. The POS tagging of Manipuri Text document follows the CRF based POS tagging as mention in [8]. In order to check the polarity of the sentiment a lexicon of verbs with manual annotation is used.

Once the verbs are identified the identified verbs have a look up to the lexicon table to identify the polarity of the sentiment. Accordingly, the polarity is notified for each the verb.

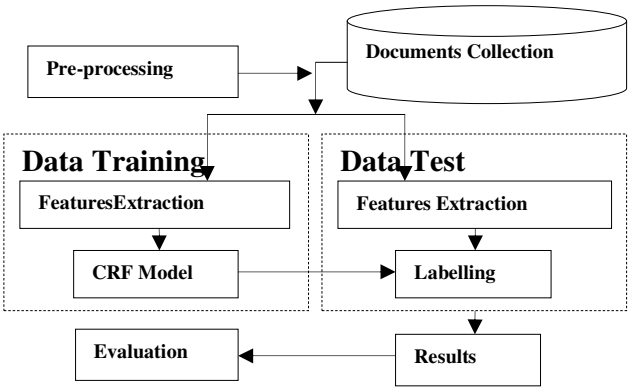

Figure 1. System block diagram

The number for each of the polarity is counted in the text document. The sentiment decider of the letter to the editors of the local daily newspapers is decided with the highest number of the polarity among the three.

A. POS tagging using CRF

As mention above the POS tagging of Manipuri text document follows the feature selection based on [8]. Fig. 2 shows the setup of the CRF based POS tagging in the Manipuri text.

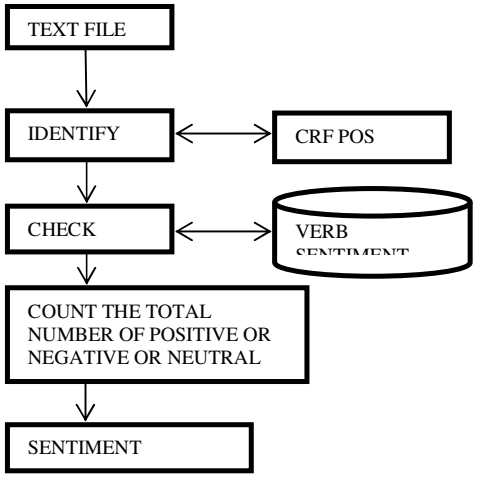

Figure 2. CRF based POS tagging

The working of CRF is mainly based on the feature selection. The feature listed for the POS tagging is as follows:

F= {W_{i-m}, … ,W_{i-1}, W_i, W_{i+1}, ..., W_{i+n}, SW_{i-m}, …, SW_{i-1}, SW_i, SW_{i+1},... , SW_{i-n} , number of acceptable standard suffixes, number of acceptable standard prefixes, acceptable suffixes present in the word, acceptable prefixes present in the word, word length, word frequency, digit feature, symbol feature, RMWE}

The details of the set of features that have been applied for POS tagging in Manipuri are as follows:

1. Surrounding words as feature: Preceeding word(s) or the successive word(s) are important in POS tagging because these words play an important role in determining the POS of the present word.
2. Surrounding Stem words as feature: The Stemming algorithm mentioned in [9] is used. The preceding and the following stemmed words of a particular word can be used as features. It is because the preceding and the following words influence the present word POS tagging.
3. Number of acceptable standard suffixes as feature: As mention in [9], Manipuri being an agglutinative language the suffixes plays an important in determining the POS of a word. For every word the number of suffixes are identified during stemming and the number of suffixes is used as a feature.
4. Number of acceptable standard prefixes as feature: Prefixes plays an important role for Manipuri language. Prefixes are identified during stemming and the prefixes are used as a feature.
5. Acceptable suffixes present as feature: The standard 61 suffixes of Manipuri which are identified is used as one feature. The maximum number of appended suffixes is reported as ten. So taking into account of such cases, for every word ten columns separated by a space are created for every suffix present in the word. A "0" notation is being used in those columns when the word consists of no acceptable suffixes.
6. Acceptable prefixes present as feature: 11 prefixes have been manually identified in Manipuri and the list of prefixes is used as one feature. For every word if the prefix is present then a column is created mentioning the prefix, otherwise the "0" notation is used.
7. Length of the word: Length of the word is set to 1 if it is greater than 3 otherwise, it is set to 0. Very short words are generally pronouns and rarely proper nouns.
8. Word frequency: A range of frequency for words in the training corpus is set: those words with frequency <100 occurrences are set the value 0, those words which occurs >=100 are set to 1. It is considered as one feature since occurrence of determiners, conjunctions and pronouns are abundant.
9. Digit features: Quantity measurement, date and monetary values are generally digits. Thus the digit feature is an important feature. A binary notation of '1' is used if the word consist of a digit else '0'.
10. Symbol feature: Symbols like $,% etc. are meaningful in textual use, so the feature is set to 1 if it is found in the token, otherwise 0. This helps to recognize Symbols and Quantifier number tags.
11. Reduplicated Multiword Expression (RMWE):(RMWE) are also considered as a feature since Manipuri is rich of RMWE. The work of RMWE is mention in [11].

5. SYSTEM DESIGN

The experiment starts with the collection and cleaning of letter to the editors from the leading local daily newspapers. The collection is of 550 letters with an average of 500 words. In total there are about 2,75,000 words.

The lexicon of verbs is collected from [10] and the polarities of each word are manually added. By polarity it means positive or negative or neutral. For positive polarity the expert marked with PV, for the negative polarity the expert marked with NG and for neutral the expert marked with NU. Sample example is shown in the Fig. 3.

The text document is run for POS tagging using CRF.The C++ based CRF++ 0.53 package[1] is used in this work and it is readily available as open source for segmenting or labeling sequential data. After this process each work is marked with the POS tag. The words which are tag with

[1] http://crfpp.sourceforge.net/

verbs are identified with a simple algorithm. These identified verbs then look up to the index of the modified lexicon of verbs. By modified verbs it means the polarity tag verbs.

The total number of negative polarity verbs is counted. Likewise the numbers of positive and neutral polarity verbs are also counted.

```
…….
…….
□□ NG
□□□ NG
□□□ NG
□□□ NG
□□□ NG
□□□ NG
……
□□. PV
□□ NU

…….
…….
```

Figure 3. Sample example of Modified Lexicon Verb List

For each of the letter to the editor document the evaluation is done with the parameter of Recall, Precision and F-score as follows:

$$\text{Recall, } \mathbf{R} = \frac{No\ of\ \ correct\ ans\ given\ by\ the\ system}{No\ of\ \ correct\ ans\ in\ the\ text}$$

$$\text{Precision, } \mathbf{P} = \frac{No\ of\ \ correct\ ans\ given\ by\ the\ system}{No\ of\ \ ans\ given\ by\ the\ system}$$

$$\text{F-score, } \mathbf{F} = \frac{(\beta^2 + 1)\ PR}{\beta^2 P + R}$$

Where β is one, precision and recall are given equal weight.

The system shows an averagerecall of 72.10%, a precision of 78.14% and a F-measure of 75.00% which is also listed in Table I.

TABLE I. AVERAGE RESULT

Model	Average Recall	Average Precision	Average F-Score
CRF	**72.10**	**78.14**	**75.00**

6. CONCLUSIONS

The work on sentiment analysis of Manipuri has to go miles. This work is just the beginning of this highly agglutinative Indian Schedule language. More methods and algorithms are to be search and implemented in order to improve the accuracy. The works can also be extended with improvement to other domains of blog, articles, twits, feedbacks, SMS etc.

REFERENCES

[1] B. Pang and L. Lee. Opinion mining and sentiment analysis.Foundations and Trends in Information Retrieval 2(1-2):1–135, 2008.

[2] Esuli, A and Sebastiani, F. 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. LREC-06

[3] Das, D. and Bandyopadhyay, S. 2010. Identifying EmotionTopic - An Unsupervised Hybrid Approach with RhetoricalStructure and Heuristic Classifier. In proceedings of the 6thIEEE-NLPKE, doi: 10.1109/NLPKE.2010.5587777

[4] A. Das and S. Bandyopadhay, "SentiWordNet for Indian languages," Asian Federation for Natural Language Processing, China, pp. 56–63, August 2010.

[5] A. Das and S. Bandyopadhay, "Subjectivity Detection in English and Bengali: A CRF-based Approach," In Proceedings of the 7th International Conference on Natural Language Processing, Macmillan 2009.

[6] J. Zhao, K. Liu and G. Wang, "Adding Redundant Features for CRFs- based Sentence Sentiment Classication" In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp.117-126, 2008

[7] Lafferty, J., McCallum, A. and Pereira, F. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proc. of the 18th International Conference on Machine Learning (ICML01).Williamstown, MA, USA. pp. 282-289

[8] Kishorjit, N. and Sivaji, B., "A Transliteration of CRF Based Manipuri POS Tagging", In the Proceedings of 2nd International Conference on Communication, Computing & Security (ICCCS-2012), Elsevier Ltd, 2012

[9] Kishorjit, N., Bishworjit, S., Romina, M., Mayekleima Chanu, Ng. & Sivaji, B., (2011) A Light Weight Manipuri Stemmer, In the Proceedings of Natioanal Conference on Indian Language Computing (NCILC), Chochin, India

[10] NingombaM. S., "A Dictionary of Manipuri Verbs", MALADES, Imphaal, 2010

[11] Kishorjit Nongmeikapam, Nonglenjaoba L., Nirmal Y. & Sivaji Bandhyopadhyay, Reduplicated MWE (RMWE) Helps in Improving the CRF Based Manipuri POS Tagger, International Journal of Information Technology Convergence and Services (IJITCS) Vol.2, No.1, DOI : 10.5121/ijitcs.2012.2106, 2012, p.45-59.

Authors

Kishorjit Nongmeikapam is working as Asst. Professor at Department of Computer Science and Engineering, MIT, Manipur University, India. He has completed his BE from PSG college of Tech., Coimbatore and has completed his ME from Jadavpur University, Kolkata, India. He is presently doing research in the area of Multiword Expression and its applications. He has so far published 30 papers and presently handling a Transliteration project funded by DST, Govt. of Manipur, India. He is the author of the Book, "See the C Programming Language".

Dilipkumar Khangembam is presently a student of Manipur Institute Of Technology. He is pursuing his B.E. in Dept. of Computer Science and Engineering. His area of interest is NLP.

Wangkheimayum Hemkumar is presently a student of Manipur Institute Of Technology. He is pursuing his B.E. in Dept. of Computer Science and Engineering. His area of interest is NLP.

Professor Sivaji Bandyopadhyay is working as a Professor since 2001 in the Computer Science and Engineering Department at Jadavpur University, Kolkata, India. His research interests include machine translation, sentiment analysis, textual entailment, question answering systems and information retrieval among others. He is currently supervising six national and international level projects in various areas of language technology. He has published a large number of journal and conference publications.

Shinghajit Khuraijam is presently working as Senior Software Enggineer at Accenture. He has Industry experience of 5 yrs, expert in Portal development and User interface. He pursued his B.E in Electronics Engg from Mumbai University and M.tech in Information Technology from CEG, Anna University, Chennai