A Karaka Based Approach to Cross Lingual Sentiment Analysis

Vartika Rai, Sakshee Vijay, and Dipti Misra Sharma

***Abstract*—Sentence level translation and projection of polarity is way more challenging for english-hindi pair as hindi is relatively free order language and translation cost adds up to the error rate. Instead of that, if we assume each sentiment bearing sentences as combination of certain karaka relations and labels, and narrow down our observation to phrase level for every sentence, there can be a dependency between english and hindi phrases through phrase level translation and finding sentiments for those particular english phrases can lead us to predicting sentiments of their respective hindi sentences as well, with very less translation errors and no dependency on hindi labelled corpus.**

***Index Terms*—Cross lingual, dependency parser, karaka roles, sentiment analysis.**

I. Introduction

Sentiment Analysis has been a topic of interest in the past decade. Humans have always been very expressive in terms of writing or giving opinions. There is data present in form of news articles, gadget reviews, editorials on political issues, tourism guides and with the growing use of social media, tweets and posts do add to it. There are a number of companies who want their company reviews to be analysed in order to get a sense of their public outlook. Users are also interested in knowing the sentiment polarity of the document before reading it any further depending on their inclination for the topic.

Hindi is one of the most commonly spoken languages in the world. Hindi is written in devanagari script. With recent advances, more amount of hindi data is present in electronic form. Hindi is a resource scarce language, where the parsers aren't that efficient and data needs to be manually tagged for every task due to the unavailability of well annotated standard corpora. The Hindi sentiwordnet available contains limited number of sentiment bearing words like adjectives, adverbs, etc.

Hindi is a free word order language which mostly uses the subject-object-verb order, whereas English uses the subject-verb-object word order. In Hindi, prepositions succeed the noun or pronouns they qualify whereas Prepositions usually come after the pronoun or noun in English. Because of these factors direct machine translation accuracy is low and hence we tried to incorporate this by direct translation of phrases rather than the whole sentence and predicting the sentiment on entire sentence level.

Karakas are an important constituent of Hindi language. Karaka relations express syntactico-semantic or semantico-syntactic relationship between verbs and nouns or pronouns in a sentence. They capture certain level of semantics closer to thematic relations. A vibhakti is assigned to each karaka, in Paninian grammar.

The structure of the paper is as follows: in the next section we describe the Literature survey. Section III introduces the dataset used for this experiment. Section IV introduces the challenges faced.. The algorithm of our system is described in Section V. Finally the last section concludes the paper.

II. Literature Survey

Ref. [1] has already created manually annotated corpus and SentiWordNet for Hindi based on equivalent of English. They learnt the best parameters for different approaches and compared their performance for sentiment analysis in Hindi.

Ref. [2] developed a lexicon of adjectives and adverbs with polarity scores using Hindi Wordnet, along with an annotated corpora of Hindi Product Reviews where they reached an accuracy of approximately 79%.

Ref. [3] improvised the existing Hindi SentiWordNet and Proposed new rules for negation handling and discourse relation for Hindi language reviews. Their Proposed algorithm produced 82.89% for positive reviews and 76.59 % for negative reviews, and an overall accuracy of 80.21%.

The process of sentiment analysis is divided into five steps [4]: Process of Sentiment Analysis for Text (Lexicon Generation), Subjectivity Detection, Sentiment polarity Detection, Sentiment Structuration, Sentiment Summarization-Visualization-Tracking.

Ref. [5] proposed system for sentiment analysis of Hindi movie review uses HindiSentiWordNet (HSWN) to find the overall sentiment associated with the document and polarity of words in the review are extracted from HSWN and then final aggregated polarity is calculated which can sum as either positive, negative or neutral. Synset replacement algorithm was used to find polarity of those words which don't have polarity associated with it in HSWN. Negation and discourse relations which are mostly present in Hindi movie review were also handled to improve the performance of the system.

In English, Philip Stones developed General Inquirer system, the first milestone for extracting textual sentiment. It was based on the manual database containing set of positive or negative orientations and the input words are compared with database to identify their class such as positive, negative, feel, pleasure [6].

Again In English, Hatzivassiloglou was the first to develop empirical method of building sentiment lexicon for

Manuscript received February 16, 2017; revised July 2017. This work was supported in part by the Kohli Center on Intelligent Systems of International Institute of Information Technology, Hyderabad.

The authors are with International Institute of Information Technology, Hyderabad, India (e-mail: vartika.rai@research.iiit.ac.in, sakshee.vijay@research.iiit.ac.in, dipti@iiit.ac.in).

doi: 10.18178/ijlll.2017.3.4.137

adjectives. The key point is based on the nature of conjunctive joining the adjectives. A log-linear regression model is provided with 82% accuracy [7].

Pang build sentiment lexicon for movie reviews to indicate positive and negative opinion. This system motivated the other machine learning approaches like Support Vector Machine, Maximum Entropy and Naive Bayes [8].

III. Dataset

The data we use are of multiple domain, crawled from various gadget, movie reviews websites such as:

http://jagran.com,
http://www.patrika.com,
http://www.bhaskar.com and

Hindi-English Parallel corpus released for ACL-Shared task (2005).

Cleaning and Pre-processing Data

After the data is crawled and is in raw text form, the next task is several iterations of processing on data, which involves:

1. Correcting the spellings in order to make it easier for further mappings to any other language (such as english), data resource(such as wordnet). In below example, the original word in corpus with its translated english equivalent is mentioned in bracket, and then , the same word after spelling correction and its correct english form is mentioned.
    1). अभीन्न (Abinn) -> अभिन्न (Integral)
    2). हॉलमार्क (H◌ॅalmark)-> हॉलमार्क (Hallmark)
2. Appending the missing end marker of sentence “|”.

IV. Challenges

The basic naive method for performing sentiment analysis in Hindi again comprises the issue of non gold datasets and the scarcity of various other resources and tools. Annotated corpus is the foremost requirement for any Machine Learning and NLP task, irrespective of purpose and requirements in predicting sentiments of corpus, be it on any level (sentence, phrase, document). A good corpus both in terms of quality and quantity plays a major role in the net performance and effectiveness of a system.

There are many datasets which have set benchmark in their quality for resource-rich languages like English and are freely available too, e.g., SemEval 2014 datasets [20]. But, Indian languages are still lagging in terms of such high quality resources. And the existing Datasets which are specific to Indian languages, created by respective research labs are very few in number with lots of limitations due to translation errors while translating a english dataset and smaller size because of it, to name a few. Examples can be seen in [1], [2], [16], [20], [21].

So, to handle this problem, we propose a cross-lingual method which involves significantly reduced translation errors and hence better coverage in terms of data and hence its sentiment values.

V. Algorithm

We use the concept of karaka roles to extract key phrases in the hindi sentence and then move on to find its English equivalent , and find the polarity of that particular English phrase and project it back to hindi. Here, we define keyphrases as combination of NN-JJ words or words that modify and give sentiment to the main verb of sentence.

For each given sentence:

We run the dependency parser on it, and obtain the output in following format:

Input : सरकार प्रतिस्पर्धा अधिकारियों को सक्रिय रहने को भी प्रोत्साहित कर रही है।

Input(English Translation) : The Government is also encouraging the competition authorities to be more proactive.

Output:

TABLE I: Karaka Roles

Index	Word	Root	POS Tag	Rel Index	Label
1	सरकार	सरकार	JJ	10	k1
2	प्रतिस्पर्धा	प्रतिस्पर्धा	NN	3	mod
3	अधिकारियों	अधिकारी	NN	6	k4
4	को	को	PSP:को	3	lwg__psp
5	सक्रिय	सक्रिय	JJ	6	pof
6	रहने	रह	VM	10	vmod
7	को	को	PSP:को	3	lwg__psp
8	भी	भी	RP	6	lwg__rp
9	प्रोत्साहित	प्रोत्साहित	JJ	10	pof
10	कर	कर	VM	0	main
11	रही	रह	VAUX	10	lwg__vaux
12	है	है	VAUX	10	lwg__vaux
13	.	.	.	10	rsvm

This table depicts relations between each word in sentence to other words using karaka roles.

1	सरकार	सरकार	JJ	10	k1

This depicts that the word सरकार (government) at index 1 is related to word at 10th position (rel index) कर, which is main verb of sentence through the label k1 (which means karta, doer of the action). Hence सरकार is the doer of the action, which is depicted by main verb.

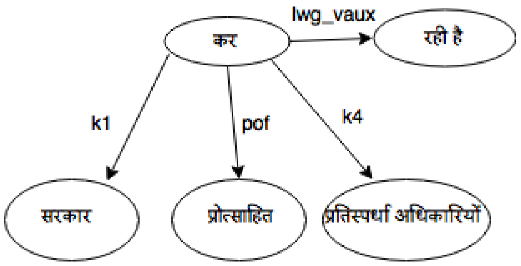
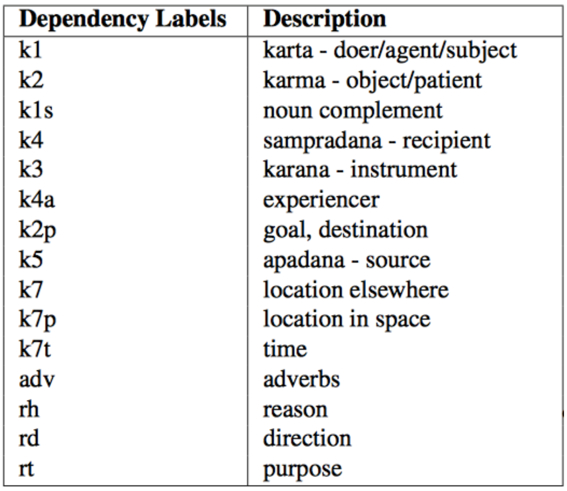


Fig. 1. Relation of main verb with other words through labels.



Dependency Labels	Description
k1	karta - doer/agent/subject
k2	karma - object/patient
k1s	noun complement
k4	sampradana - recipient
k3	karana - instrument
k4a	experiencer
k2p	goal, destination
k5	apadana - source
k7	location elsewhere
k7p	location in space
k7t	time
adv	adverbs
rh	reason
rd	direction
rt	purpose

Fig. 2, All Dependency Labels in Hindi Treebank

By the help of these marked relations of each word with verb, we extract various properties of sentence, which will be useful for us, namely:

The agent of action: k1

The action performed: k2

Doer of action: k3

Part of relations: pof

Modifier for Nouns: nmod / nmod_adj

Approach 1: Coarse Grain Classification (Sentence Level)

For this, we extract phrases from this output on following rules:

For each relation which involves the dependency label as

-pof (Part of relation, Part of units such as conjunct verbs)

-mod/nmod/nmod_adj (all kinds of modifier to main verb/ verb modifiers(vmod))

Extract those relations and make small phrases out of them, with given assumption that these convey and hold maximum sentiment bearing parts of sentence (hence Keyphrases of sentence).

According to above rule, the phrases which we obtain are:

प्रोत्साहित कर रही है

प्रतिस्पर्धा अधिकारियों को सक्रिय रहने

After these phrases are obtained, we conduct phrase level translation and translate these phrases into English:

प्रोत्साहित कर रही है

Is Encouraging

प्रतिस्पर्धा अधिकारियों को सक्रिय रहने

After we find the english equivalent of hindi key phrases, to obtain overall sentiment pattern of sentence as positive or negative, we cross reference and find weighted label of adjective and noun in the english phrase from MPQA corpus, with weights given according to type of word as strong/weak subject.

Adjective/Noun in phrases: encouraging, active

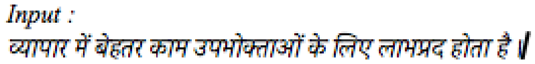
type=strongsubj len=1 word1=encouraging pos1=adj stemmed1=n priorpolarity=positive

type=weaksubj len=1 word1=active pos1=adj stemmed1=n priorpolarity=positive

Hence, it leads to overall sentiment of sentence as positive which is correct.

Approach 2 (Using ESWN to sum polarities):

Due to limitation of the words in it, and to avoid conflict situations while predicting sentence classification is tough due to multiple subjects with different polarity, i.e if the sentence sums up to having two weak subject negative polarity words and one strong subject positive polarity words, one strong negative polarity words and three weak positive polarity words etc. , we come up with an approach where we predict the polarity of translated english phrases by mapping it to ESWN (English sentiwordnet), hence attaching polarity score of every word and performing weighted average of phrase score to come up with overall polarity score of sentence.



Input :

व्यापार में बेहतर काम उपभोक्ताओं के लिए लाभप्रद होता है ।

Input (English Translation): And better performance in business in turn benefits consumers.

Output:

TABLE II: KARAKA LABELS

Index	Word	Root	POS Tag	Rel Index	Label
1	व्यापार	व्यापार	NN	9	k7
2	में	में	PSP :में	1	lwg__psp
3	बेहतर	बेहतर	JJ	4	nmod__adj
4	काम	काम	NN	9	k1
5	उपभोक्ताओं	उपभोक्ता	NN	9	rt
6	के	का	PSP:का	5	lwg__psp
7	लिए	ले	PSP:ले	5	lwg__psp
8	लाभप्रद	लाभप्रद	JJ	9	pof
9	होता	हो	VM	0	main
10	है	है	VAUX	10	lwg__vaux
11	.	.	.	10	rsvm

Phrases obtained by previously defined rules and their scores:

लाभप्रद होता है: 0.625

बेहतर काम होता है: 0.5

Overall score: average of phrase polarities: 0.56, hence positive sentence , which is correct as per analysis.

VI. CONCLUSION

The above method uses labels generated from dependency parsers to indicate relations between each word in a sentence, and then utilise these labels to perform cross lingual sentiment prediction with least translation errors as phrase translations have less error rate and greater accuracy as compared to entire sentence translation. Also, since the method is not dependent on various available hindi labelled corpuses and datasets, the coverage is significantly better. We can extend this work to fine grain level where we can predict the sentiment score/label with respect to each topic in the sentence. Also, this approach can be adopted to supervised method in which trained phrases can be used to predict sentiment labels.

ACKNOWLEDGMENT

The research conducted in this project has been supported by Kohli Center on Intelligent Systems (KCIS) at International Institute of Information Technology, Hyderabad (IIITH), India. We also thank them for making available dependency parser developed at IIIT freely downloadable to NLP researchers. We have used the following link: https://bitbucket.org/sivareddyg/hindi-dependency-parser.

REFERENCES

[1] J. Aditya, A. R. Balamurali, and P. Bhattacharyya. "A fall-back strategy for sentiment analysis in hindi: A case study," in *Proc. the 8th ICON*, 2010.

[2] B. Akshat, P. Arora, and V. Varma. "Hindi subjective lexicon: A lexical resource for Hindi polarity classification," in *Proc. Eight International Conference on Language Resources and Evaluation (LREC)*, 2012.

[3] M. Namita *et al.*, "Sentiment analysis of Hindi review based on negation and discourse relation," in *Proc. International Joint Conference on Natural Language Processing*, 2013.

[4] K. Amandeep and V. Gupta, "A survey on sentiment analysis and opinion mining techniques," *Journal of Emerging Technologies in Web Intelligence*, vol. 5, no. 4, 2013, pp. 367-371.

[5] P. Pooja and S. Govilkar. "A framework for sentiment analysis in Hindi using HSWN," *International Journal of Computer Applications*, vol. 119, no. 19, 2015.

[6] J. S. Philip, D. C. Dunphy, and M. S. Smith, "The general inquirer: A computer approach to content analysis," 1966.

[7] H. Vasileios and K. R. McKeown. "Predicting the semantic orientation of adjectives," in *Proc. Eighth Conference on European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, 1997.

[8] W. Janyce *et al*., "Learning subjective language," *Computational Linguistics*, vol. 30, no. 3, 2004, pp. 277-308.

[9] D. Amitava and S. Bandyopadhyay, "SentiWordNet for Indian languages," *Asian Federation for Natural Language Processing*, China, 2010, pp. 56-63.

[10] P. Bo, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proc. the ACL-02 Conference on Empirical Methods in Natural Language Processing*, vol. 10, 2002.

[11] D. Amitava, and S. Bandyopadhyay, "Subjectivity detection in English and Bengali: A crf-based approach," in *Proc. ICON*, 2009.

[12] A. M. Shad, A. Ekbal, and P. Bhattacharyya, "Aspect based sentiment analysis in Hindi: Resource creation and evaluation," in *Proc. the 10th edition of the Language Resources and Evaluation Conference (LREC)*, 2016.

[13] S. Richa, S. Nigam, and R. Jain, "Polarity detection movie reviews in Hindi language," *arXiv Preprint arXiv: 1409.3942* (2014).

[14] B. Naman and U. Z. Ahmed, "Advisor: Amitabha mukherjee," *Sentiment Analysis in Hindi, Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur, India*, 2013, pp. 1-10.

[15] T. Pranali, A. S. Sambare, and S. R. Jain, "Opinion mining in natural language processing using Sentiwordnet and fuzzy."

[16] A. R. Balamurali, A. Joshi, and P. Bhattacharyya, "Robust sense-based sentiment classification," in *Proc. the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, Association for Computational Linguistics, 2011.

[17] B. Rafiya *et al*., "Dependency annotation scheme for Indian languages," *IJCNLP*. 2008.

[18] N. Joakim, J. Hall, and J. Nilsson. "Maltparser: A data-driven parser-generator for dependency parsing," in *Proc. LREC*, vol. 6, 2006.

[19] H. Samar *et al*., "The ICON-2010 tools contest on Indian language dependency parsing," in *Proc. ICON-2010 Tools Contest on Indian Language Dependency Parsing, ICON* 10, 2010, pp. 1-8.

[20] P. Maria *et al.*, "Semeval-2014 task 4: Aspect based sentiment analysis," in *Proc. SemEval*, 2014, pp. 27-35.

[21] A. R. Balamurali, "Cross-lingual sentiment analysis for Indian languages using linked wordnets, 2012.

[22] M. Sneha, "Sentiment classification in Hindi," *International Journal of Scientific and Technology Research*, vol. 3, no. 5, 2014.

[23] K. Ayush *et al*., "IIT-TUDA: System for sentiment analysis in Indian languages using lexical acquisition," *International Conference on Mining Intelligence and Knowledge Exploration*, Springer International Publishing, 2015.

[24] J. Sarthak and S. Batra. "Cross lingual sentiment analysis using modified BRAE," *EMNLP*, 2015.

[25] Y. Mukesh and V. Bhojane. "Sentiment analysis on Hindi content: A survey," *International Journal of Innovations and Advancement in Computer Science (IJIACS)*, 2015.

[26] A. R. Kashyap, P. Balamurali, P. Bhattacharyya, and G. Haffari, "The haves and the have-nots: Leveraging unlabelled corpora for sentiment analysis," 2013.

[27] D. Erkin, "Cross-lingual sentiment analysis with machine translation," M.S. thesis. 2013

[28] A. B. Ram, *Hindi Dependency Parsing and Treebank Validation*, Diss. International Institute of Information Technology Hyderabad, 2011.

[29] S. Richard *et al.*, "Recursive deep models for semantic compositionality over a sentiment Treebank," in *Proc. the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, vol. 1631, 2013.

Vartika Rai was born in Varanasi, U.P. on 17 July 1994. She completed her BTech in computer science and is currently pursuing MS by Research in Computational Linguistics in the Kohli Center on Intelligent Systems under the guidance of Prof. Dipti Mishra Sharma at International Institute of Information Technology, Hyderabad.

Sakshee Vijay was born in Jaipur, U.P. on 22 July 1994. She completed her BTech in computer science and is currently pursuing MS by Research in Computational Linguistics in the Kohli Center on Intelligent Systems under the guidance of Prof. Dipti Mishra Sharma at International Institute of Information Technology, Hyderabad.

Dipti Misra Sharma is the head of the Language Technologies Research Center and a professor at International Institute of Information Technology, Hyderabad (IIIT-H), India. Dr. Sharma did her post-graduation and Ph.D. degree in linguistics from University of Delhi. After spending a year in Germany working on a project on Orality in written literatures', she came back and first joined Osmania University and then moved to University of Hyderabad (UoH) as a UGC research associate. Thereafter, she worked as a faculty at UoH teaching various courses in applied linguistics.