Opinion Mining System for Processing Hindi Text for Home Remedies Domain

Arpana Prasad
Department of Computer Science, Punjabi University, Patiala.
arpanaprasad2013@gmail.com

Neeraj Sharma
Department of Computer Science, Punjabi University, Patiala.
sharma_neeraj@hotmail.com

Shubhangi Sharma
SAP Labs India Ltd.
shubhangi.sharma@sap.com

Abstract

Lexical and computational components developed for an Opinion Mining System that process Hindi text taken from weblogs are presented in the paper. Text chosen for processing are the ones demonstrating cause and effect relationship between related entities 'Food' and 'Health Issues'. The work is novel and lexical resources developed are useful in the current research and may be of importance for future research.

1 Introduction

Opinion Mining (OM) is a field of study in Computer Science that deals with development of software applications related to text classifications and summarizations. Researchers working in this field contribute lexical resources, computing methodologies, text classification approaches, and summarization modules to perform OM tasks across various domains and different languages. The common challenges encountered by the researchers in this field are; domain dependency, classification of sarcastic text, dealing with comments with thwarted expectations, language dependency, extensive domain knowledge, and categorising of text as a fact bearing text or an opinion bearing fact. There has been development in this field for processing Hindi unstructured text(Arora, 2013; Arora, Bakliwal, & Varma, 2012; Harikrishna & Rao, 2019; Jha, Savitha, Shenoy, Venugopal, & Sangaiah, 2017; Joshi, R, & Bhattacharyya, 2010; Mishra, Venugopalan, & Gupta, 2016; Mogadala & Varma, 2012; Reddy & Sharoff, 2011; Sharma, Nigam, & Jain, 2014) .

2 Research Problem

An ongoing research in the field of OM, which is dedicated towards the development of lexical and software components that facilitate opinion classification and summarization from Hindi Text appearing on Web logs is presented here. Hindi text showcasing cause and effect relationship between 'Food' and 'Health Issues' entities are processed in the OMS under study. The resources are developed for an algorithm 'A' such that for a sentence 'Y' which is a domain specific sentence from weblogs in Hindi, A(Y) returns a set {F, HI, p, s} such that F is a subset of set, FOOD={set of word or phrases in Hindi used for an edible item and HI is a subset of set, HEALTH_ISSUE= {{set of word or phrases in Hindi used for a part of body composition 'BODY_COMPONENT'} UNION {set of word or phrases in Hindi used for a health problem a human being face 'HEALTH_PROBLEM'}}. Element 'p' takes numeric value '1' or '-1' where value '1' means that from the text 'Y', algorithm 'A' computationally derived that the food entities mentioned in set 'F' have a positive effect in health issues mentioned in set 'HI' and the numeric value '-1' means that the food entities in set 'F' have a negative effect in health issues in set 'HI'. The element's' may take value '1' or '2' indicating that the strength of polarity 'p' is medium or strong. This research is undertaken with a sole objective to contribute towards bringing text appearing in Hindi weblogs under the benefits that an OM system offers. Language dependent computational contributions based on English and Chinese languages, in the previous studies motivated the proposal of the current work (Bhattacharya, 2014; Miao, Zhang, Zhang, & Yu, 2012; Yang, Swaminathan, Sharma, Ketkar, & D'Silva, 2011). This research is undertaken with a sole objective to contribute towards bringing text appearing in Hindi

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 21–23
Patna, India, December 18 - 21, 2020.

weblogs under the benefits that an OM system offers.

3 Key Issues Identified and Addressed

Lexical resources required for Named Entity Recognition (NER) and polarity classification were are not available for the current research. The same are developed in the research. A corpus in same domain and in same language as the text being processed in an OM system helps in devising computational components for the system and evaluating them. An annotated corpus of Hindi text relevant to the research did not pre-exist hence it is developed in the research.

4 Major Contributions

A domain specific Hindi corpus, with semantic and syntactic annotations is developed in the research. The semantic annotations of the corpus help in devising and evaluating the algorithms formulated in the research. A total of 3303 unique domain specific Hindi sentences from weblogs are collected for the corpus. The corpus has approximately 2516 unique sentences with positive annotations for 'FOOD' and 'HEALTH ISSUE' associations and 787 unique sentences with negative annotations for associations between entities. The total number of syntactically annotated Hindi words in the corpus is 60000 approximately. Domain specific lexical datasets for entities; FOOD, FOOD_ADJECTIVE, FOOD_COMPONENT, BODY_COMPONENT and HEALTH_PROBLEM are also developed/identified in the research. The total number of Hindi words/phrases identified for the lexical datasets are 8000 lexicons approximately. A lexical based polarity classification and polarity strength classification algorithm is developed in the research. There are two datasets developed to support the algorithm. The datasets are; (a) a set of approx. 15000 positive polarity bearing phrases and a set of approx. 10000 negative polarity bearing phrases that generate a vocabulary of approximately 35000 trigram words, (b) a set of approximately 14000 strength determining phrases that are useful in determining the strength of the polarity identified by the algorithm.

5 Methodology Adopted

Firstly, a set of domain specific keywords were consolidated. Then for approximately 2 years using some identified keywords domain specific Hindi text from weblogs were collected. Semi automated processes were adopted for data cleaning and refinement of the collected sentences for the corpus. The corpus is syntactically annotated with part of speech (POS) using a POS tool for Hindi developed by CFILT, IIT Bombay. Each word of the sentence is stemmed to give flexibility to the word usage for NER. The corpus is semantically annotated with the help of two annotators. A detailed guideline for semantic annotations was developed in the research. The annotators were supposed to annotate the text on their perception about the related entities and polarity of associations of the related entities. Domain specific lexical datasets are developed using Hindi WordNet of CFILT, IIT Bombay. The developed lexical datasets are used by the NER algorithm that extracts domain specific related entities from text under OM processing. The polarity bearing and strength determining phrases are developed using a seed list from the corpus, later the dataset was extended using phrases from books that are candidate for appearing on weblogs and using synonyms of words in phrases with same POS using Hindi WordNet. Finally, a lexical based algorithm using Naive Bayes Classifier that is trained on a vocabulary of n-gram words from polarity phrases and strength determining phrases is formalized for polarity classification of association and strength classification of polarity.

6. Experimental Results

All the lexical resources are developed using SQLite Studio 3.1.1. The algorithms are developed using Python 3.7. The classification algorithms when trained using trigram words from polarity bearing phrases and strength determining phrases and tested on a random set of 900 sentences from the corpus gives best results with Accuracy: 0.996, Precision: 0.998 Recall: 0.994 and F-Score of 0.996. The named entity recognition algorithm tested on the same dataset gives 85% accuracy. The lexical outcomes of this research may be useful to other researchers working in related fields.

References

Arora, P. (2013). Sentiment Analysis for Hindi Language. International Institute of Information Technology. Retrieved from coling2017.pdf (iiit.ac.in)

Arora, P., Bakliwal, A., & Varma, V. (2012). Hindi Subjective Lexicon Generation using WordNet Graph Traversal. International Journal of Computational Linguistics and Applications, 3(Jan-Jun 2012), 25–29.

Bhattacharya, S. (2014). Computational methods for mining health communications in web 2. 0. University of Iowa. Retrieved from http://ir.uiowa.edu/etd/4576

Harikrishna, D. M., & Rao, K. S. (2019). Children's Story Classification in Indian Languages Using Linguistic and Keyword-based Features. ACM Transactions on Asian and Low-Resource Language Information Processing, 19(2). https://doi.org/https://doi.org/10.1145/3342356

Jha, V., Savitha, R., Shenoy, P. D., Venugopal, K. R., & Sangaiah, A. K. (2017). A novel sentiment aware dictionary for multi-domain sentiment classification, 0, 1–13. https://doi.org/10.1016/j.compeleceng.2017.10.015

Joshi, A., R, B. A., & Bhattacharyya, P. (2010). A Fall-back Strategy for Sentiment Analysis in Hindi: a Case Study. In Proceedings of the ICON-2010:8th International Conference on Natural Language Processing. Macmillan Publishers, India. Retrieved from http://ltrc.iiit.ac.in/proceedings/ICON-2010

Miao, Q., Zhang, S., Zhang, B., & Yu, H. (2012). Extracting and Visualizing Semantic Relationships from Chinese Biomedical Text. Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation, 99–107. Retrieved from http://aclweb.org/anthology/Y12-1010

Mishra, D., Venugopalan, M., & Gupta, D. (2016). Context Specific Lexicon for Hindi Reviews. Procedia Computer Science, 93(September), 554–563. https://doi.org/10.1016/j.procs.2016.07.283

Mogadala, A., & Varma, V. (2012). Retrieval approach to extract opinions about people from resource scarce language news articles. Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining - WISDOM '12, (August), 1–8. https://doi.org/10.1145/2346676.2346680

Reddy, S., & Sharoff, S. (2011). Cross language POS taggers (and other tools) for Indian languages: An experiment with Kannada using Telugu resources. Cross Lingual Information Access, 11–19. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.222.1557&rep=rep1&type=pdf#page=27

Sharma, R., Nigam, S., & Jain, R. (2014). Polarity Detection of Movie Reviews in Hindi Language. International Journal on Computational Science & Applications, 4(4), 49–57. https://doi.org/10.5121/ijcsa.2014.4405

Yang, H., Swaminathan, R., Sharma, A., Ketkar, V., & D'Silva, J. (2011). Mining biomedical text towards building a quantitative food-disease-gene network. Studies in Computational Intelligence, 375, 205–225. https://doi.org/10.1007/978-3-642-22913-8_10arXiv:1503.06733. Version 2