A Comprehensive Analysis of Guided Abstractive Text Summarization

Jagadish S Kallimani[1], Srinivasa K G[2] and Eswara Reddy B[3]

**[1] Research Scholar, Department of Computer Science and Engineering
Kakinada-533003, Andhra Pradesh, India**

**[2] Professor, Department of Computer Science and Engineering
Bangalore-560054, Karnataka, India**

**[3] Professor, Department of Computer Science and Engineering
Ananthapuramu-515002, Andhra Pradesh, India**

Abstract

Abstractive summarization is the process of creating a condensed version of the given text document by collating only the important information in it. It also involves structuring the information into sentences that are simple and easy to understand. This paper presents the process that generates an abstractive summary by focusing on a unified model with attribute based Information Extraction (IE) rules and class based templates. Term Frequency/Inverse Document Frequency (TF/IDF) rules are used for classification of the document into several categories. Lexical analysis reduces prolixity, resulting in robust IE rules. Usage of templates for sentence generation makes the summaries information intensive. The IE rules are designed to accommodate the complexities of some Indian languages. This paper analyzes the adaptation of the system methodology over multiple Indian languages and several document categories. It also draws comparison between abstracts generated and summaries obtained by extractive methods.

Keywords: *Template based Generation, Language Parsing and Understanding, Information Extraction, Template Selection, Abstractive and Extractive Text Summarizations.*

1. Introduction

The field of abstractive summarization, despite the rapid progress in Natural Language Processing (NLP) techniques, is a persisting research topic. Even in global languages like English, the present abstractive summarization techniques are not all quintessential due to shortcomings in semantic representation, inference and natural language generation [1]. Research in abstractive summarization methodologies for Indian regional languages has been started upon only recently. It became more challenging due to lack of linguistic tools in abstractive summarization for Indian languages.

Earlier research on summarizing documents in Indian languages adopted paradigms for extracting salient sentences from text using features like word frequency and phrase frequency [2], position in the text [3] and key phrases [4]. Such extractive summaries tend to have long sentences and the desired information is scattered across the document. Extractive summarization does not combine, either syntactically or semantically, concepts mentioned in the different text spans of the source document. In contrast, abstractive approaches like sentence compression, sentence fusion [5], which are rewriting techniques based on syntactic analysis, are currently being explored. But these methods provide little improvement over extractive methods in terms of content selection.

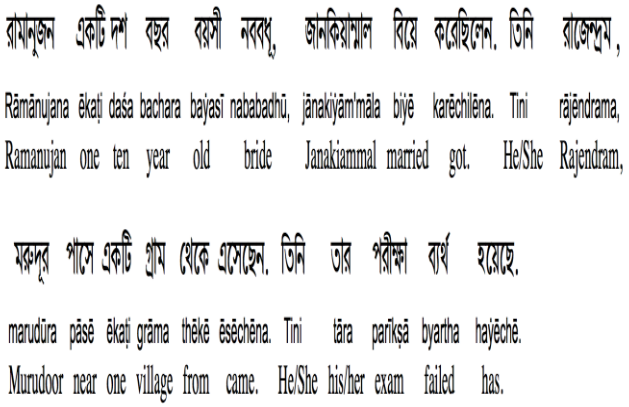
The methodology proposed in this paper, uses topic based guided summarization technique [4] to present a concise abstractive summary. Initiating with *Kannada* language, the concept has been established with languages like *Hindi*, *Bengali* and *Telugu* each with their own set of complexities.

2. Indian Languages

The languages of India belong to several language families, the major ones being the Indo-Aryan languages (*Hindi, Bengali, Gujarati, Marathi, and Punjabi*) and the Dravidian languages (*Tamil, Telugu, Malayalam, Kannada, and Tulu*). On a global scale, Kannada alone has 35.5 million speakers around the world, which is greater than the population of Australia (which stands at 23.4 million) [6], while Hindi, Bengali, Telugu have 366 million, 207 million, 69.7 million speakers respectively [7]. These statistics further motivate the need of a summarizer for Indian languages.

Both these styles, Indo Aryan and Dravidian - have much in common; They lack letter case distinction and words tend to be highly inflected i.e., information that relates a verb to its noun is expressed by means of post-positions of nouns (called *"vibhaktis"*). Hence, the treatment of inflection, not word order, plays the most important role in processing unlike English, where word order has a central role.

The inflection of words (*vibhakti*) in Indian regional languages indicates the place of the event, the victim, etc. Also, stylistic variations, agglutination, differences in gender treatment between languages are some of the complexities to be handled while analyzing the text.



রামানুজন একটি দশ বছর বয়সী নববধূ, জানকিয়াম্মাল বিয়ে করেছিলেন. তিনি রাজেন্দ্রম,

Rāmānujana ēkaṭi daśa bachara baẏasī nababadhū, jānakiẏām'māla biẏē karēchilēna. Tini rājēndrama,

Ramanujan one ten year old bride Janakiammal married got. He/She Rajendram,

মরুদূর পাসে একটি গ্রাম থেকে এসেছেন. তিনি তার পরীক্ষা ব্যর্থ হয়েছে.

marudūra pāsē ēkaṭi grāma thēkē ēsēchēna. Tini tāra parīkṣā byartha haẏēchē.

Murudoor near one village from came. He/She his/her exam failed has.

Fig. 1. Usage of neutral gender in *Bengali* language

Figure 1 illustrates complexity in Bengali. In Bengali, gender is neutral and there is no special separation for male and female. Adjectives are also usually not modified according to the number or case of the nouns they qualify. In the above example, it is difficult to identify who is from '*Rajendra*' village and who failed in the examination.

Previously, attempts were made to generate extractive text summary for Kannada and Telugu languages by the same authors [8] [9]. Related articles such as [10] and [11] are also referred for support vector machine and texture feature extraction based on wavelet transforms respectively.

3. Motivation

With the increasing need for automatic summarizers in the context of data mining and NLP, there is huge scope for research work and this study is an attempt to achieve it. It is an intuitive choice as the language of study with an impressive online presence when compared to English and other global languages.

4. Problem Statement

This work aims to blend IE methods to guided summarization of text documents by using tagging rules like Named Entity Recognition (NER). The generation of abstract summary of the document has been proposed for the study. The present study deals with the following objectives:

- To develop an abstractive content-aware summary of single documents of text.
- To ensure retrieval of content relevant to aspects of each category of documents considered.
- To develop a method of forming different sentences to present the information extracted.
- To produce simple, easy to understand and cohesive text, that conveys important aspects of the original text document.

5. The Methodology

A fully abstractive approach with a separate process for the analysis of the text, the content selection, and the generation of the summary promises to have the most potential for generating summaries that are comparable to that written by humans [12]. The basic flow of the system is represented in Figure 2.

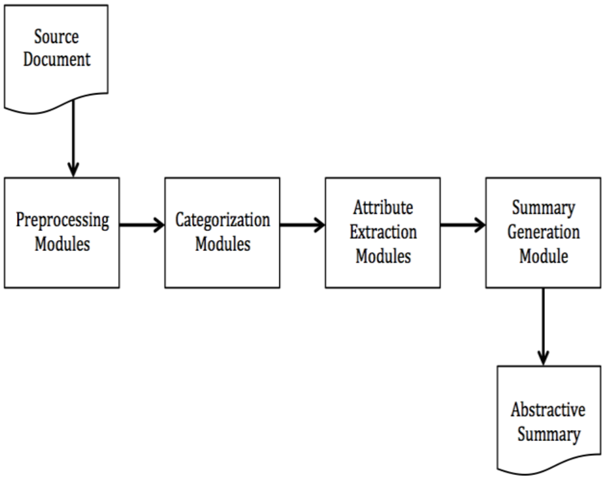


Fig. 2. Overview of the system

5.1 Preprocessing and Categorization

The input document is subjected to pre-processing like lemmatization and stemming along with Parts of Speech (POS) tagging using a cross lingual tool [13]. Identification of named entities like names, locations and dates also form an integral part of this phase. Repositories of rules and gazetteers are complied to assist entity identification in the NER phase.

Categorization serves as an indication of the context of information that is to be included in the summary generated by the system. The document needs to be classified into a specific category based on the information to be extracted. A TF-IDF rule based classifier is used for this purpose. After eliminating stop words and their derivatives, frequency of the terms and their relevance in that document is determined. This set of frequent terms is compared against category specific keywords which determine the category of the document.

Once the document is categorized, the information or attribute extraction modules associated with that category are applied on the document.

5.2 Attributes and Classes

The methodology interprets highly predictable elements called attributes, which follow a guide called class, presenting a very specific, unified information model of the given topic.

Attributes are category specific, primary pieces of information that are expected to be present in the summary.

Classes are configurations or blueprints, which indicate the multiple attributes to be identified for a given topic. The IE rules extract candidate answers for these attributes. Multiple classes can be merged to handle documents belonging to more than one category.

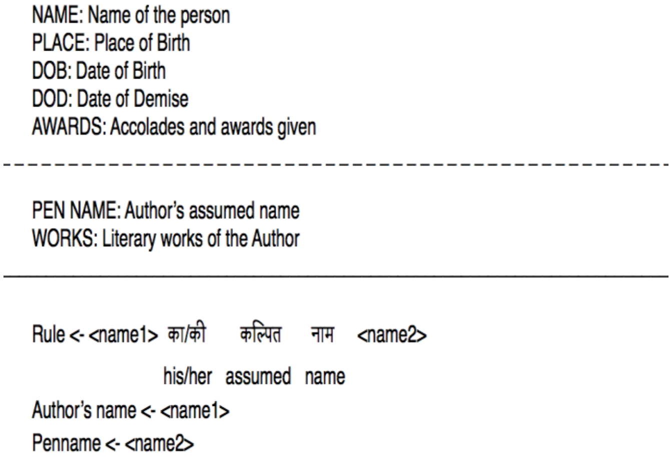
The input document is mainly a free-text document. It could be a news report, or a description of a phone handset, or biography of some eminent personality.

The document to be summarized is subjected to preprocessing, namely – POS Tagging and NER. Also, TF/IDF rule based classifier is used to categorize the document, which in turn determines the classes to be applied to it.

The POS and NE tags along with synonymous verb and noun forms help in crafting the IE rules in identifying the roles of interest. Gazetteers and noun inflections are used to extract the required information by including them in the IE rules.

Figure 3 gives an example of a class for the category Biography. Apart from general attributes applicable to all biographies, specialized class for literary personalities can be combined with the generic Biography class to cover documents in that field. The class also has multiple IE rules to extract the attributes. In the example above, the rule tries to extract the pen name of the author.
Redundant information may be captured by the IE modules which only serve to reinforce the validity and salience of the attribute chosen. Several content selection heuristics are employed and a compact information piece for summary generation is chosen.

The key pieces of information expected in the summary for a given topic remains static and hence, no reconstruction of attributes is required irrespective of the language of the document. Only the IE rules need to be inflexed to accommodate the coils of other languages.



NAME: Name of the person
PLACE: Place of Birth
DOB: Date of Birth
DOD: Date of Demise
AWARDS: Accolades and awards given

PEN NAME: Author's assumed name
WORKS: Literary works of the Author

Rule <- <name1> का/की कल्पित नाम <name2>
his/her assumed name
Author's name <- <name1>
Penname <- <name2>

Fig. 3. Example of a Class for the category *Biography*

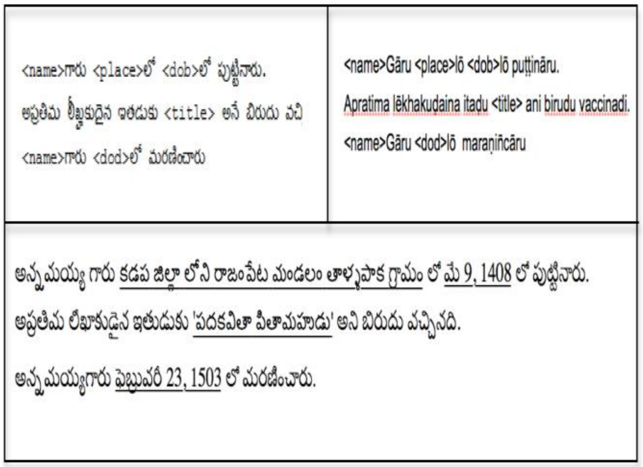
5.3 Template based sentence generation

The final step in realizing the summary is filling up the template. Templates are natural language generating systems that map their non- linguistic input directly (i.e., without intermediate representations) to the linguistic surface structure. Templates are generic structures of sentences with crucial pieces of information missing. The attributes extracted in the previous stage are mapped to this text plan, to deliver the information in an effective manner. Template based sentence generation, though conceptually straightforward, necessitates the extracted information piece to be compounded with the right inflections. This impels root word extraction and appropriate transformations on the attributes to facilitate the completion of sentences in the template.

A possible drawback of using template based sentence generation is monotony in the sentence structure of the generated summaries. A comprehensive set of templates help in generating variety of sentence formations and information deliverables based on the category.

This novel IE rule-based approach attempts to extract relevant information using lexical analysis tools like POS tagging and NER. This ensures an information rich summary that reduces redundancy in not just the sentences produced but also in the information conveyed. The algorithm is as follows:

1. *Perform POS Tagging and Stemming on input text document.*
2. *Recognise named entities like person, locations, dates, etc using gazetteers and rules.*
    2.1. *Identify category of the text document using statistical methods like TF.*
    2.2. *Extract information for Aspects of the corresponding scheme using IE rules.*
3. *Select appropriate template and populate it to generate a summary.*



<name>గారు <place>లో <dob>లో పుట్టినారు.
అప్రతిమ లేఖకుడైన ఇతడుకు <title> అనే బిరుదు వచి
<name>గారు <dod>లో మరణించారు

<name>Gāru <place>lō <dob>lō puṭṭināru.
Apratima lēkhakuḍaina itaḍu <title> ani birudu vaccinadi.
<name>Gāru <dod>lō maraṇiñcāru

అన్నమయ్య గారు కడప జిల్లా లోని రాజంపేట మండలం తాళ్ళపాక గ్రామం లో మే 9, 1408 లో పుట్టినారు.
అప్రతిమ లేఖాకుడైన ఇతుడుకు 'పదకవితా పితామహుడు' అని బిరుదు వచ్చినది.
అన్నమయ్యగారు ఫెబ్రువరి 23, 1503 లో మరణించారు.

Fig. 4. Sample template in Telugu and the summary generated

Figure 4 shows a sample template and the generated summary for Biography category in *Telugu* language. The placeholders given in angular brackets indicate the position of the different attributes to be replaced in the template. The underlined text in the summary indicates the attributes extracted from the document.

Although the system depends on domain knowledge, shallow NLP and hand-written IE rules, it can easily be expanded to cover a plethora of focus groups. The system makes a clear distinction between the NLP stage and just extraction of relevant Information from a given text. This allows a seamless integration of new information extraction rules and methods with existing ones. The current implementation designs classes to be reusable to many topics.

6. Results and Discussion

6.1 Intrinsic Evaluation

For many NLP problems, the definition of a gold standard is a complex task, and can prove impossible when inter-annotator agreement is insufficient. The unavailability of a single golden standard that systems should conform to is one of the main challenges in automatic text summarization. Usually, manual evaluation is performed by human judges, who are instructed to estimate the quality of a system, or most often of a sample of its output, based on a number of criteria [14].
The Precision and recall are computed as:

$$\text{precision} = \frac{\text{number of relevant items retrieved}}{\text{total number of items retrieved}}$$

$$\text{recall} = \frac{\text{number of relevant items retrieved}}{\text{number of relevant items in collection}}$$

The current system proposes an equal weightage to precision and recall of the summary produced (Beta = 1). Another relevant score is accuracy, which is measured as:

Accuracy= (number of true positives + number of true negatives)/ (number of true positives + false positives + false negatives + true negatives)

Assume an IR system has recall R and precision P on a test document collection and an information need. The F-measure of the system is defined as the weighted harmonic mean of its precision and recall, that is,

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

Where the weight $\alpha \in [0, 1]$. The balanced F-measure, commonly denoted as F_1 or just F, equally weighs precision and recall, which means $\alpha = 1/2$. The F_1 measure can be written as

$$F_1 = \frac{2PR}{P+R}$$

This table shows the evaluation of the system over six different categories. The average Precision, Recall, Accuracy, F-Measure values for each category is tabulated.

Documents across categories like Biographies, Natural Calamities, Product reviews, Cultural Events, Cricket, and Attacks were considered as initial case study.

An intrinsic evaluation was performed where fifteen human judges were given the task of identifying attributes in each category. The resulting schemas were used as golden standard for evaluation against the system. The number of attributes recognized for a category remains fairly constant irrespective of the length of the input document.

The system achieved an average 86.24% precision, 78.93% recall, and 81.5% F-measure across six different categories as shown in the table 1.

Table 1
Evaluation of the proposed abstractive system

Category	Precision	Recall	Accuracy	F measure
Biographies	0.9	0.73	0.77	0.81
Natural Calamities	1	0.92	0.92	0.96
Product Reviews	1	0.91	0.91	0.95
Cultural Events	0.71	1	0.71	0.83
Cricket	0.793	0.676	0.575	0.73
Attacks	0.7714	0.5	0.43	0.61
Average	**0.8624**	**0.7893**	**0.7191**	**0.815**

Figure 5 shows the recall-precision graph for over 30 documents in different categories. The system maintains good precision and recall values on the whole. The average precision histogram measures the average precision of a run on each topic against the median average precision of all corresponding runs on the topic.

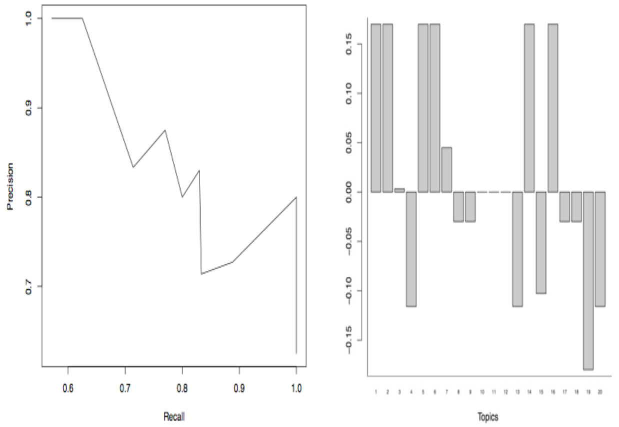


Fig. 5. Recall-Precision graph and Average Precision Histogram

6.2 Extrinsic Evaluation

The non-existence of a generic abstractive text summarization system for Indian languages has made the evaluation of summaries generated by the system abstruse.

Thus, as an extrinsic evaluation, comparison against summaries generated by an extractive summarizer was conducted. The emphasis in extraction methods is usually on determining salient text units (typically sentences) by looking at the text unit's lexical and statistical relevance or by matching phrasal patterns [15]. The extractive summarizer considered for comparison uses General Social Survey (GSS) coefficients and IDF methods along with TF for extracting key words and uses them to generate the summary.

A comparison of the summary generated by the system and that of the extractive summary is conducted. The evaluation was performed over three different categories namely Biography, Cricket and Natural Calamity. The extract is limited to the top ten highly ranked sentences and is compared with the abstractive summary which has ten sentences in it.

Table 2
Evaluation of Extractive Summarizer

Category	Precision	Recall	Accuracy	F measure
Biography	0.27	0.6	0.44	0.375
Cricket	0.8	1	0.8	0.88
Natural Calamity	0.38	1	0.466	0.55
Average	**0.4833**	**0.866**	**0.5686**	**0.6016**

This table shows evaluation results of the extractive summarizer over three categories. Table 3 shows the comparison between the proposed abstractive summary evaluations with existing extractive summary evaluations. As expected, the extractive summarizer has high recall but compromises on precision.

Table 3
Abstractive Vs Extractive Test Summarization

Type	Precision	Recall	Accuracy	F measure
Abstractive	0.8624	0.7893	0.7191	0.815
Extractive	0.4833	0.866	0.5686	0.6016

In comparison with the extractive summarizer, the proposed system clearly performs better over the three categories as illustrated in table 3. It also indicates the benefits of abstractive summary in terms of crispness, information coverage, compression ratio and readability over an extractive system. The abstractive summaries also has better information coverage and compression ratio.

7. Conclusion and future work

The paper provides the methodology to create abstractive summaries of text documents written in Indian regional languages. As demonstrated by the summary generated, the methodology has proved to have good precision values apart from maintaining readability. The problem of text summarization is modeled as an IR problem. Making a clear distinction between the IE stage and NLG stage allows the addition of new classes, IE rules, and improvements in each stage without affecting the other. The challenges posed by Indian languages are handled at each stage of writing IE rules and creating generic templates.

Though instrumental in achieving the desired results, usage of templates can bring in flatness or monotony to the summaries generated. An ideal solution to break this monotony would be the use of tools like WordNet [16], which is a freely available lexical database, or SimpleNLG [17], which is a Java API designed to facilitate the generation of natural language.

Extension to speech outputs of summaries is an aspect that can be explored as part of future work. A possible variation of the system can be to produce the summary output in a more universal language like English.

References

[1] M. Kumar, D. Das, A. I. Rudnicky. Summarizing Non-textual Events with a 'Briefing' Focus. In Proceedings ofRecherche d'Information Assistée par Ordinateur (RIAO), Pittsburgh, USA, May 30-Jun 1, 2007..

[2] Jayashree R, Srikanta Murthy K and Sunny K. Keyword Extraction Based Summarization of Categorized Kannada Text Documents. International Journal on Soft Computing (IJSC), Nov 2011, Vol. 2, No. 4.

[3] Kamal Sarkar, Bengali text summarization by sentence extraction. In Proceedings of International Conference on Business and Information Management(ICBIM), NIT Durgapur, 2012, pp. 233-245.

[4] Varsha R. Embar, Surabhi R. Deshpande, Vaishnavi A. K, Vishakha Jain, Jagadish S. Kallimani. sArAmsha - A Kannada Abstractive Summarizer. In Proceedings of International Conference on Advances in Computing, Communications and Informatics (ICACCI), Mysore, India, 22-25 Aug, 2013.

[5] Amitava Das and Sivaji Bandyopadhyay. Syntactic Sentence Fusion Techniques for Bengali. In Proceedings of International Journal of Computer Science and Information Technologies, Vol. 2, Issue. 1, 2011, pp.494-503.

[6] Wikimedia Foundation, Inc. Web. List of countries by population. Wikipedia, the Free Encyclopedia. May 2014.

[7] Wikimedia Foundation, Inc. Web. List of languages by number of native speakers in India. Wikipedia, The Free Encyclopedia. May 2014.

[8] Jagadish S Kallimani, Srinivasa K G, Eswara Reddy B, Information Retrieval by Text Summarization for an Indian Regional Language, 6th International Conference on Natural Language Processing and Knowledge Engineering, IEEE NLP-KE 2010, August 21-23, 2010, Beijing, China, IEEE Catalog Number: CFP10811-PRT, ISBN:978-1-4244-6897-3, pp. 596-599.

[9] Jagadish S Kallimani, Srinivasa K G and Eswara Reddy B, Information Extraction by an Abstractive Text Summarization for an Indian Regional Language, The 7th International Conference on Natural Language Processing and Knowledge Engineering, IEEE NLP-KE 2011, Tokushima, Japan. Nov 27-29, 2011.

[10] Erhu Zhang, Fan Wang, Yongchao Li, Xiaonan Bai, Automatic detection of micro calcifications using mathematical morphology and a support vector machine, Bio-Medical Materials and Engineering. 2013, Volume 24, issue 1, 53-59,

[11] Shan Hu, Chao Xu, WeiQiao Guan, Yong Tang, Yana Liu, Texture feature extraction based on wavelet transforms and gray-level co-occurrence matrices applied to osteosarcoma diagnosis, Bio-Medical Materials and Engineering. 2013. Volume 24, issue 1, 129-143.

[12] Genest, Pierre-Etienne, and Guy Lapalme. Text generation for abstractive summarization. In Proceedings of the Third Text Analysis Conference. Gaithersburg, Maryland, USA. National Institute of Standards and Technology. 2010.

[13] Siva Reddy and Serge Sharoff. Cross Language POS Taggers (and other Tools) for Indian Languages: An Experiment with Kannada using Telugu Resources. In Proceedings of IJCNLP workshop on Cross Lingual Information Access: Computational Linguistics and the Information Need of Multilingual Societies. Chiang Mai, Thailand. 2011.

[14] John Dragomir R. Radev, Eduard Hovy, Kathleen McKeown. Introduction to the special issue on summarization. Association for Computational Linguistics, Vol. 28, No. 4. 2002. doi: http://dx.doi.org/10.1162/089120102762671927

[15] U. Bruce Hahn, Mani I. The challenges of automatic summarization. IEEE-Computer, 33(11), 29–36, 2000. DOI:http://dx.doi.org/10.1109/2.881692

[16] George A. Miller, WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11, 39-41, 1995, doi:http://dx.doi.org/10.1145/219717.219748

[17] Gatt and E. Reiter. SimpleNLG: A realization engine for practical applications. In Proceedings of ENLG, 2009.

Jagadish S Kallimani is pursuing his research in the areas of information extraction, text summarization, speech synthesis and prosody analysis for several Indian regional languages. He has published many papers in national and international conferences and journals in the above areas. He is currently interested in developing hand written rules for word sense disambiguation and CFG for summarization.

Srinivasa K G received his PhD in Computer Science and Engineering from Bangalore University in 2007. He is now working as a professor and Head in the Department of Computer Science and Engineering, M S Ramaiah Institute of Technology, Bangalore. He is the recipient of All India Council for Technical Education - Career Award for Young Teachers, Indian Society of Technical Education – ISGITS National Award for Best Research Work Done by Young Teachers, Institution of Engineers(India) – IEI Young Engineer Award in Computer Engineering, Rajarambapu Patil National Award for Promising Engineering Teacher Award from ISTE - 2012, IMS Singapore – Visiting Scientist Fellowship Award. He has published more than hundred research papers in International Conferences and Journals. He has visited many Universities abroad as a visiting researcher – He has visited University of Oklahoma, USA, Iowa State University, USA, Hong Kong University, Korean University, National University of Singapore are few prominent visits. He has authored two books namely File Structures using C++ by TMH and Soft Computer for Data Mining Applications LNAI Series – Springer. He has been awarded BOYSCAST Fellowship by DST, for conducting collaborative Research with Clouds Laboratory in University of Melbourne in the area of Cloud Computing. He is the principal Investigator for many funded projects from UGC, DRDO, and DST. His research areas include Data Mining, Machine Learning and Cloud Computing.

Dr. B. Eswara Reddy Graduated in B.Tech (CSE) from Sri Krishna Devaraya University in 1995. He received Masters Degree in M.Tech (Software Engineering), from JNT University, Hyderabad, in 1999. He received Ph.D in Computer Science & Engineering from JNT University, Hyderabad, in 2008. He served as Assistant Professor from 1996 to 2006, Associate professor from 2006 to 2012.He served as Head of the Dept. of CSE during 2010 to 2012. Presently, he has been serving as professor of CSE Dept and Coordinator for Master of Science in Information Technology (MSIT) program, JNTUA (in collaboration with CIHL and CMU). He has been acting as Board of Studies (BoS) – UG Chairman since 2011 to till date for JNT University Anantapur.
He has more than 30 Publications in various International Journals and more than 20 Publications in various National and International Conferences. He has authored two books: “Programming with Java published by Pearson/Sanguine Publishers” and “Data Mining: Principles and approaches” published. He has received University Grants Commission-Major Research Project(UGC-MRP) titled ‘Cloud computing framework for rural health care in Indian scenario’ His research interests include Cloud Computing, Pattern Recognition & Image Analysis, Data Warehousing & Mining and Software Engineering. He is a life member of ISTE, IE, ISCA, IAENG and member of CSI, IEEE.

.