Abs-Sum-Kan: An Abstractive Text Summarization Technique for an Indian Regional Language by Induction of Tagging Rules

Shilpa G V, Shashi Kumar D R

ABSTRACT--- This paper presents a full abstraction for Indian languages, specifically Kannada, in the context of guided summarization. The proposed process generates the abstractive sum-mary by focusing on a unified presentation model with aspect based Information Extrac-tion (IE) rules and scheme based Templates. TF/IDF rules are used for classification into categories. Lexical analysis (like Parts Of Speech tagging and Named Entity Recognition) reduces prolixity, which leads to robust IE rules. Usage of Templates for sentence genera-tion makes the summaries succinct and information intensive. The IE rules are designed to accommodate the complexities of the considered languages. Later, the system aims to produce a guided summary of domain specific documents. An abstraction scheme is a collection of aspects and associated IE rules. Each abstraction scheme is designed based on a theme or subcategory. An extensive statistical and qualitative evaluation of the summaries generated by the system has been conducted and the results are found to be very promising.

Keywords— Abstractive Summary, Information Extraction, Kannada, Template based Generation, Template Selection.

1. INTRODUCTION

Abstractive summarization is the process of creating condensed version of a given text document by collating only the important information in it. Also it involves structuring them into sentences that are simple and easy to understand. The non-linear nature of Indian languages and lack of linguistic tools make abstractive summarization a daunting task.

The World Wide Web contains a multitude of documents and is growing at an exponential pace. Major sources of information on the web like Wikipedia offer Kannada versions of its Wiki pages. Thus the increasing availability of information online has created the problem of information overload. This has led to extensive research in the field of automatic text summarization in Natural Language Processing (NLP). Summarization is also fundamental to many other NLP and data mining applications such as information retrieval, text clustering and so on [1].

Kannada language is spoken in India predominantly in the state of Karnataka, making it the 25th most spoken language in the world. There is also a considerable difference between the spoken and written forms of this language. Spoken Kannada tends to vary from region to region. The written form is more or less consistent throughout Karnataka. It is the official and administrative language of Karnataka [2]. Based on the recommendations of the Committee of Linguistic Experts appointed by the Ministry of Culture, the Government of India officially recognised Kannada as a classical language [3].

The versatility of the method has been expounded by applying it to a miscellany of languages. Initiating with Kannada, the concept has been established with languages like Hindi, Bengali and Telugu each with their own different complexities.

2. MOTIVATION

With the increasing need for automatic summarizers in the context of Data Mining and NLP, there is tremendous scope for research work and this study is an attempt to achieve it. Creating an abstractive summary especially for a regional language like Kannada is challenging. It is an intuitive choice as the language of study with an impressive online presence when compared to English and other global languages.

3. PROBLEM STATEMENT

This work aims to tailor IE methods to guided summarization of Kannada text documents by using tagging rules like NER. The development of a system which accepts text documents and generates an abstract summary of the document has been proposed by considering Kannada as the language of study. The present study deals with the following objectives:

- To develop an abstractive content-aware summary of single documents of Kannada 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.

4. EARLIER WORKS

Automatic text summarization [4] has been in existence since 1950s. It has been actively researched in recent years and several automatic summarization methods have been

Revised Manuscript Received on July 10, 2019.

Shilpa G V, Vemana Institute of Technology, Bangalore. Karnataka,India.(shilpa.gv@vemanait.edu.in)

Shashi Kumar D R, Cambridge Institute of Technology, Bangalore. Karnataka,India. (shashikumar.cse@citech.edu.in)

proposed. Symbolic techniques using parsers, grammars, and semantic representations, do not scale up to real-world size. IR and other statistical techniques, being based on word counting and word clustering, cannot create true summaries because they operate at the word (surface) level instead of at the concept level. Text Summarization [5] can be divided into three steps - Topic Identification or extraction, Interpretation, and Generation.

There are mainly two approaches to summarization – extraction and abstraction. Extractive summaries [1] are formulated by extracting key text segments (sentences or passages) from the text, based on statistical analysis of individual or mixed surface level features to locate the sentences to be extracted. Extractive systems have become statistically indistinguishable in evaluation results [6]. The "most important" content is treated as the "most frequent" or the "most favourably positioned" content. Such an approach thus avoids any efforts on deep text understanding. They are conceptually simple and easy to implement.

In Abstractive summarization [7], attempts have been made to develop an understanding of the main concepts in a document and then express those concepts in clear natural language. It uses linguistic methods to examine and interpret the text.

Earliest instances of research on summarizing documents proposed paradigms for extracting salient sentences from text using features like word frequency and phrase frequency, position in the text and key phrases. Computationally, features of sentences that are useful to score sentences for potential inclusion in the summaries have been proposed. The [8] shows similar frequency extraction and set of occurrence between each term and the frequent terms. MEAD [9] extends this idea and computes the score of a sentence based on many surface level features such as similarity to the first sentence of the document, position of the sentence in the document, sentence length etc. More recently, machine learning approaches such as neural networks [10], support vector machines have been applied to carry out text categorization and summarization [11]. The neural network method involves training the neural networks to learn the types of sentences that should be included in the summary. Three phases are generally involved – training, pruning, and feature fusion. Similar approaches have also been adopted for Kannada texts [12].

In contrast, Abstraction is generally considered more powerful than extraction, giving a concise and succinct summary. As they are based on a formal representation of the document's content, they adapt well to high compression rates, such as those needed for wireless Personal Digital Assistants (PDAs) and similar technologies. Thus, abstractive summarization has not been researched to the same extent as extractive summarization. Sentence compression [13], sentence fusion [14] and sentence splitting [15], are few existing approaches. All are rewriting techniques based on syntactical analysis, offering little improvement over extractive methods in the content selection process. Opinosis proposes a novel flexible summarization framework that uses graphs to produce abstractive summaries of highly redundant opinions. In contrast with the previous work, Opinosis assumes no domain knowledge and uses shallow NLP, leveraging mostly the word order in the existing text and its inherent redundancies to generate informative abstractive summaries.

Unsupervised document summarization method creates the summary by clustering and extracting sentences from the original document [16]. A comparative study on abstractive summarization methods is conducted based on text representation, content selection and summary generation [17]. The information from the source text is extracted into the form of abstract data which is post processed to infer the most important message from the original text [18].

In this paper, the methodology proposed relies on information extraction and domain based templates to create an abstractive summary.

4.1 Part Of Speech Tagging and NER

Linguistic approaches involve analysis of the text (Part of Speech Tagging - POST, Named Entity Recognition (NER), Ontology based terminology extraction, etc.), content selection and sentence generation. POST is the process of marking up a word in a text as corresponding to a particular part of speech, based on both its definition, as well as its context. POS tagging has been explored extensively in Kannada [19] [20].

Named Entity Recognition (NER) is used to locate and classify atomic elements in text into predetermined classes such as the names of persons, organizations, locations, concepts etc. Named Entity became a popular term in NLP and was introduced in the sixth Message Understanding Conference (MUC-6) [21].

It is essential to recognize information units like names, including person, organization and location names, and numeric expressions including time, date, money and percent expressions. NER and NLP research around the world has progressed in leaps and bounds in the last two decades with the advent of effective machine learning algorithms, creation of large annotated corpora for various languages and using generative statistical models like HMM (Hidden Markov Model).

Research is not extensive in resource poor Indian languages like Kannada. Recently, statistical discriminative models like Condition Random Fields (CRF) [22] and labelling sequence data are used consistently for segmenting and labelling the sequence data as a graphical model [23]. The impact of textual genre (journalistic, scientific, informal, etc.) and domain (gardening, sports, business, etc.) has been rather neglected in the NER literature. Few studies are specifically devoted to diverse genres and domains [24]. It is a system designed for emails, scientific texts and religious texts. In [25], a system is created specifically designed for email documents.

4.2 Real vs. Template based Natural Language Generation (NLG)

NLG is the process of constructing outputs from non-linguistic inputs. An NLG system is a translator that converts a computer based representation into a natural language representation. A simpler system, SimpleNLG [26], requires Java programming knowledge. This

knowledge cannot be assumed for content and subject matter experts. Thus a generic method is required.

But [27] argues that template-based systems can, in principle, perform all NLG tasks in a linguistically well founded way. Many implemented systems of this kind deviate dramatically from the stereotypical systems that are often associated with the term template.

Keeping in mind the limited availability of Kannada resources and complexities involved in the language, a feasible methodology is proposed.

5. CHARACTERISTICS OF KANNADA

The language uses forty-nine phonemic letters, divided into three groups: *swaragalu* (vowels – thirteen letters); *vyanjanagalu* (consonants – thirty-four letters); and *yogavaahagalu* (neither vowel nor consonant – two letters: the *anusvara* ◌ಂ and the *visarga* ◌ಃ).

Each of the Kannada characters is represented using UNICODE. The character set is almost identical to other Indian languages. The script is complicated by the occurrence of various combinations of "half-letters" (glyphs), or symbols that attach to various letters (*samyuktaksharas*). Each written symbol in the Kannada script corresponds with one syllable, as opposed to one phoneme in languages like English. The Kannada script is syllabic.

The characteristics and difficulties that are associated with Kannada language are:

- No Capitalization
- Non-availability of large gazetteer
- Lack of standardization and spelling, e.g. ಕಾರ್ಯ {karya}[1] [work][2] can also be written as ಕಾರ್ಯಾ {karya}[1] [work][2] .
- Number of frequently used words (common nouns) which can also be used as names are very large. Also the frequency with which they can be used as common noun as against person name is more or less unpredictable.

eg: ಆನಂದ {Anandha} [ananda] (‘name of a person’ as well as ‘being happy’)

- Lack of annotated corpora
- Scarcity of resources and tools
- Free word order language

eg: ನಾನು ಆನೇ ತರಗತಿಯಲ್ಲಿ ಒದುತ್ತಿದ್ದೇನೆ. {nanu 8ne taragathiyalli odhuthiddhene} [I am studying in 8th standard]

ಆನೇ ತರಗತಿಯಲ್ಲಿ ನಾನು ಒದುತ್ತಿದ್ದೇನೆ. {8ne taragathiyalli nanu odhuthiddhene} [In 8th standard, I am studying]

- Ambiguity in Parts of Speech

e.g. ಕರಿ {kari} [black] can mean:

Meaning	POS Tag
Black	Adjective
To Fry	Verb
Elephant	Noun

- Words change spellings when the stems are inflected

E.g.: ದೊಡ್ಡಾಸ್ಪತ್ರೆ {dhoDDaspatre} [big hospital] -> ದೊಡ್ಡ + ಆಸ್ಪತ್ರೆ {dhoDDa+Aspatre} [big+hospital]

ಬೆಂಗಳೂರಿನಲ್ಲಿ {bengaLurinalli} [in Bangalore] -> ಬೆಂಗಳೂರು + ಅಲ್ಲಿ {bengaLuru+alli} [Bangalore+in]

Compound bases, called *samāsa* in Kannada, are a set of two or more words compounded together.

Examples: ತಂಗಾಳಿ {tangaaLi} [cold weather], ಹೆಮ್ಮರ {hemmara} [big tree], ಇಮ್ಮಡಿ {immaDi} [double quantity].

The Kannada script is almost perfectly phonetic, but for the sound of a "half *n*" (which becomes a half *m*). The number of written symbols, however, is far more than the forty-nine characters in the alphabet, because different characters can be combined to form compound characters (*ottakshara*).

POS tagging has been explored quite extensively in Kannada using [28] and [29]. Cross-linguistic approach was attempted [30] using Telugu (another Indian language) resources asserting that Kannada and Telugu are similarly structured. IE, NER and Word Sense Disambiguation problems have been discussed in [30].

ಬಿರ್ಲಾ	ಮೈದಾನದಲ್ಲಿ	ಗಾಂಧಿಯವರನ್ನು	ನಾಥೂರಾಮ್	ಗೋಡ್ಸೆ	ಗುಂಡಿಕ್ಕಿ	ಕೊಂಡನು.
Birlā	maidānadalli	gāndhiyavarannu	nāthūrām	gōḍse	gunḍikki	kondanu.
Birla	Ground (in)	Gandhi (was)	Nathuram	Godse	a bullet (shot)	killed (by).

The inflection of words in Kannada language is shown above. The *vibhakti “alli” [in]* associated with the word *“maidaana” (ground)* indicates the place of the event. Similarly, the victim (in this case, Gandhi) can be identified by the *vibhakti “annu” [to]*.

Text summarization has also been attempted using keyword extraction methods in [31]. They use two feature selection techniques for obtaining features from documents. Scores obtained by GSS (Galavotti, Sebastiani, and Simi) coefficients and IDF (Inverse Document Frequency) methods along with TF (Term Frequency) for extracting key words, later use these for summarization based on rank of the sentence.

6. PROPOSED ARCHITECTURE

IE is a method of filtering large text documents. This paper explains a method which combines IE with summarization to produce a guided summary of domain-specific documents. This method uses a narrower view which is to identify instances of a particular class of events and extract arguments relevant to this class of events.

A fully abstractive approach with a separate process for the analysis of the text, the content selection, and the generation of the summary has the potential for generating summaries at a level comparable to humans. The proposed method uses a rule-based, custom-designed IE module, along with categorization, content selection and sentence generation systems to fulfil the needs of abstractive summarization. The system uses repositories like rules and gazetteers to refer to the language syntax and semantics. Figure 1 shows the system design.

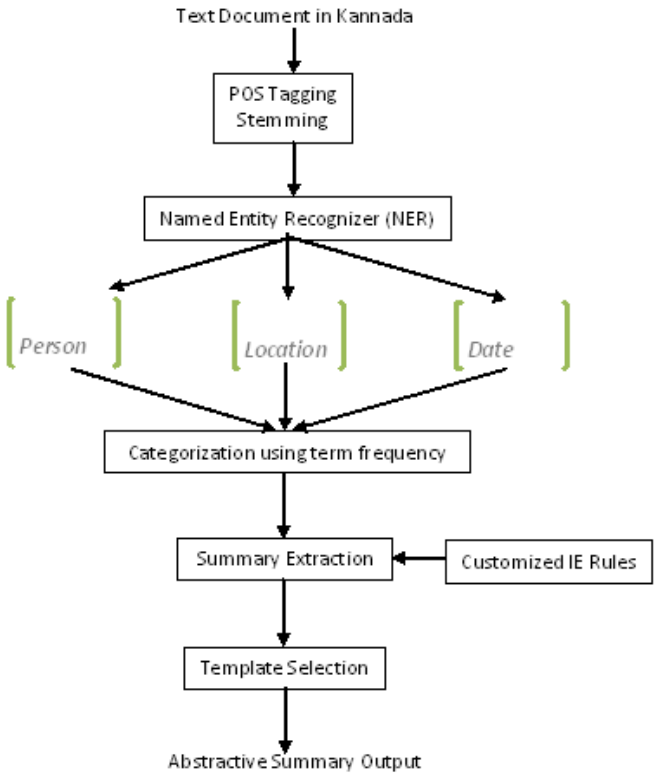

Fig. 1. Proposed Architecture for Abstractive Summary Generation System

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.*

The Kannada documents are first pre-processed and tokenized. Lemmatization and stemming is done along with POS tagging using a cross-lingual tool [32]. Identification of names, locations and dates is essential and is done by indication with a collection of tags such as BPER, IPER which give beginning of the name and its continuation. Locations and dates follow the same principle as the names. Special handwritten rules resolve ambiguity between names and locations and increase identification accuracy. Due to corpora being limited for Indian languages, gazetteers are compiled to assist entity identification [33].

Abstraction schemes are relied upon to extract information from the documents. A scheme is specific to a theme or a domain. Each scheme has a list of aspects which define the most important and essential information items that must be present in the summary generated. They also have a set of IE rules which translate pre-processing annotations to candidate answers for a specific aspect. One or more schemes can be merged to handle a single category. For example:

What: What Happened?
When: Date and Time
Where: Location

In the subcategory of Cultural Events, more detailed contextual set of aspects are added as:

Who: Performing artist
School: Institution and teacher
Entry fee: Cost of tickets

Similarly, for a document categorized as Earthquake, the same set of Aspects as in Events can be identified along with more specific aspects such as:

Victims: People affected
Property damaged: Infrastructure affected
Richter scale: Magnitude
Epicentre: Origin

The IE rules use several synonymous verb and noun forms, identifying the syntactical position of roles

of interest or aspects. The POS tags and named entities identified in the previous phase assist in crafting these rules. Another approach for rule writing makes use of gazetteers. For example, in the category of Biographies, gazetteers help to identify the prominent awards won by an artist in his field.

The various categories considered in this work include Biographies, Cricket, Natural Disaster, Bomb Blasts, Cultural Events, and Tech Reviews. A particular scheme is chosen from the scheme set based on text categorization. The strategy is to generate a list of important words in the document and compare their relevance to a category. This classification is approached statistically using TF as the frequency of occurrence of a particular word in the text. The most frequent words thus obtained, are compared with their relevance to a specific category. To increase the accuracy of the classifier, stop words and punctuations are ignored. An abstractive summary must be clear and concise apart from being grammatically correct.

As discussed earlier, Template based approach is more suitable in the scenario of the current project. The line between standard NLG systems and template based approach has become blurred as some systems combine standard NLG with templates [31]. Because modern template-based systems tend to use syntactically structured templates, and allow the gaps in them to be filled recursively (i.e., by filling a gap, a new gap may result). Some template-based systems use grammars to aid linguistic realization. In the present work, we follow the template based approach to map the information obtained in the previous phase to deliver the concise information.

A comprehensive set of templates are created which offer a variety of sentence formations and information deliverables based on the category and data that needs to be included in the summary.

Many templates with different sentence structures and formations with varying order are created to reduce monotony in the sentences generated. An example set of templates that can be used to convey the aspects in Biographies is shown in Figure 2.

<name> ಒಬ್ಬ ಅಪ್ರತಿಮ <profession>. ಇವರು <dob> ಜನಿಸಿದರು. ಹುಟ್ಟೂರು <native>. <name> is one extraordinary <profession>. He was born on <dob>. Birth place is <native>.	<name>ರವರು (ಜನನ: <dob>) <native> ಇಂದ ಬಂದ <profession>. <name>,who was born on <dob> has come from <native> <profession>.	ಪ್ರಖ್ಯಾತ <profession> ರಾದ <name>ರವರು, <dob>ರಲ್ಲಿ <native> ಅಲ್ಲಿ ಜನಿಸಿದರು. The famous <profession> <name> was born on <dob> at <native>.

Fig. 2. Sample template sentences

The generated summary should have a logical flow of information. Thus, ordering of schemes needs to be considered. To illustrate this generic nature of the approach, a small experiment using Telugu documents was also considered. The input considered is a tagged document which is fed to the system where similar extraction rules identify the aspects and generate a short summary based on templates shown in Figure 3.

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

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

Fig. 3. Sample template for Biography category in Telugu. The placeholders given in angular brackets indicate the different Aspects to be extracted and replaced in the template.

7. RESULTS

The system limits the range of information to be included in the summary in contrast to full text understanding. This helps in eliminating redundancy and expressing the most important information from the source document. The highly domain specific and aspect oriented nature of this method is advantageous in meeting the needs of user focus groups. For example, Journalists can look at the summaries for certain aspects to determine the reliability of the information and whether it is worth following up with the original source.

As proof of concept, around 50 documents were considered as initial case study. These documents belonged to various categories where each was interpreted as schemes with their own list of aspects.

The summaries are short and have high density of information. The elements of information are stated in a way that is easy to understand. Irrespective of the length of the input document, the number of aspects considered for a particular category will be constant as they are tailor made to suit specific requirements.

Since summarization is not a deterministic problem, the evaluation tends to be objective in nature. Intrinsic evaluations are normally employed, either by soliciting human judgments on the goodness and utility of a given summary, or by a comparison of the summary with a human summary.

Evaluation of a summary can also be done using metrics like Compression Ratio and Retention.

CR is computed for each automated summary generated over the original text.

CR = (length of Summary) / (length of Original Text)

Table 1. Compression Ratio

Document	Length of input document (D) in words	Length of Summary (S) in words	Compression Ratio (CR) =S/D
Doc1	547	49	0.0895
Doc2	310	51	0.1645
Doc3	205	39	0.1902
Doc4	418	59	0.1411
Doc5	366	43	0.1174

As observed, the system achieves good compression ratios, which is a significant improvement over those obtained by extractive summaries.

7.1 Aspect based Evaluation

For the evaluation of the current system, an aspect based evaluation is conducted across domains by analyzing multiple manual human abstracts similar to Summary Content Units (SCU). But due to the very nature of the summarization task, different people would choose different information and even the same person may choose different information at different times. This shows differences between summaries created by humans but the key elements will always feature in the summaries.

Thus, a manual evaluation of the summarizer against a human summary across documents is done.

- Five people who had no prior knowledge of the system were given three input documents.
- They were made to mark sentences they considered to hold important information that should be present in a summary of the same.
- The same documents were then fed to the system.
- The information covered was used as a metric to evaluate the different summaries.

Table 2 shows a comparison of the results obtained between the human summaries (Hμ) and system generated summary (Sμ). The values in the table are a ratio of Key information items (Ki), in human summary to system generated summary given by Hμ/ Sμ.

Table 2. Comparison of the results

	Human 1(H1)	Human 2(H2)	Human 3(H3)	Human 4(H4)	Human 5(H5)
Doc1	8/6	7/6	9/6	9/6	8/6
Doc2	10/6	7/6	10/6	11/6	9/6
Doc3	11/8	7/8	12/8	11/8	13/8

Table 3 shows the degree of commonality (Cμ) between the aspects considered by humans and the system.

Table 3. Cμ: Ratio of (Sμ∩Hμ) to Sμ.

	H1	H2	H3	H4	H5
Doc1	6/6	6/6	5/6	4/6	6/6
Doc2	6/6	4/6	6/6	5/6	5/6
Doc3	8/8	5/8	7/8	7/8	7/8

A variation of the above test was carried out for other domains like Calamity, Attacks, and Cricket. Human test subjects were instructed to list the aspects in the summary to cover for every category. This was then compared to the aspects covered in the system.

The results are consolidated in the bar graphs as shown in Figure 4. The vertical axis represents the number of aspects considered. The graphs compare Human Aspects and System Aspects and also indicate the number of common Aspects between the two.

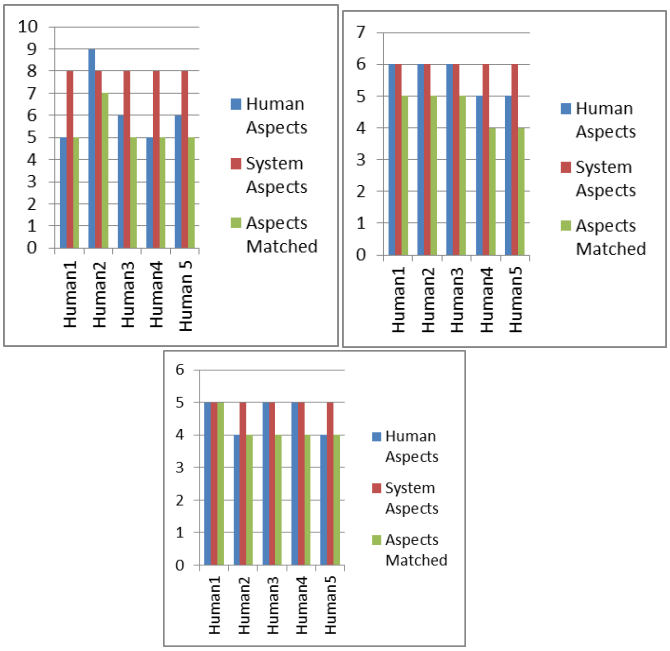

Fig. 4. Aspect comparisons for Cricket, Attacks and Natural Disaster

These results indicate a good retention of key information that is expected to be present in a summary. The human generated summaries also contained the same information items or SCUs that the system has generated. In some cases, there were additional Aspects included by some humans and in others, the Sμ exceeded Hμ. This is expected due to restrictions in length of summary to be generated and the resulting tradeoffs made between retention and Compression ratio. But in totality, the number of common aspects indicates that relevant content was delivered as intended.

7.2 Linguistic Quality Evaluation

All the evaluation methods discussed so far have been focused on evaluating the information content of a

summary. The summary readability is also an important factor in summary evaluation.

In DUC, a set of questions was developed to evaluate readability aspects of summaries. Are they ungrammatical? Do they contain redundant information? While much progress has been seen in improving system content selection, most automatic summaries score poorly on readability aspects such as coherence. Human assessments of linguistic quality on a scale, usually 1–5, are probably the fastest and cheapest to obtain. They do not require the collection of standard summaries, nor any annotation or manual analysis from the assessors. Because of these properties, this evaluation approach is rather attractive.

A Mean Opinion Score (MOS) [34] test was conducted with 15 people as our initial test field users.

- A questionnaire was prepared which consisted of queries to rate the features expected in a good summarizer. The questionnaire considered metrics to rate the reliability of the system in generating good quality summaries, ease of use and understandability and usefulness of the system among others.
- Fifteen people were asked to use the system and asked to evaluate the same.
- The rating was done on a scale of 1-5 with 5 indicating highest score and 1 being the lowest.

The test results are very optimistic and indicate that the system generated summaries are of good linguistic quality with greater satisfaction as shown below in table 4.

Table 4. MOS questionnaire for the Evaluation of the system

Users	Length: Adequate?	Is the summary simple?	Summary: Covers important information	Summary: Redundant information	Aesthetic appeal	GUI : Ease of use	Experience Satisfaction	Is the System useful?
1	5	5	4	5	4.5	4	5	5
2	4	5	4	5	5	4.5	5	4.5
3	4	5	5	5	5	5	5	4
4	5	5	4.5	5	5	4.5	5	5
5	4.5	5	5	5	4.5	5	5	5
6	5	5	5	5	5	5	4.5	5
7	3	5	5	5	5	5	4.5	5
8	5	5	5	5	4.5	5	5	5
9	4	5	5	5	4.5	5	4.5	5
10	4	5	5	5	4.5	5	5	4.5
11	4	5	4.5	4.5	4.5	5	5	4.5
12	5	5	4	5	4.5	5	5	4.5
13	4.5	5	4.5	5	5	5	5	5
14	3.75	4.75	4.75	5	3.75	5	4.75	5
15	5	4	4	5	3.75	4	4	5

The system was extensively tested for Kannada text documents as input. Further, to illustrate the generic nature of the approach, an experiment using Telugu documents was also considered. As seen in Figure 5, a given Telugu document describing the life of Saint *Annamacharya* can be summarized by mapping the rules written for Kannada biographies to Telugu script and semantics.

అన్నమయ్యగారు వైఎస్ఆర్ జిల్లా రాజంపేట మండలం తాళ్ళపాక గ్రామ లో మే9,1408లో పుట్టినారు.
అప్రతిమ లీఖ్ఖుదైన ఇతడుకు 'పదకవితా పితామహుడు ' అనే బిరుదు వచ్చినది.
అన్నమయ్యగారు ఫిబ్రవరి23,1503లో మరణించారు

Fig. 5. Sample Telugu Summary

Dravidian languages share strong features with the Indo-Aryan languages, which have been attributed to a substratum influence from them. Thus, a similar trial was conducted for documents in Hindi and Bengali (other Indian languages) by writing rules in accordance with the structure of the language.

7.3 Objective Test Results

In the presented work, the evaluation of the summary is obtained by compression ratio, retention, summary content units, comparison between human summaries and system generated summaries, and aspect comparisons. Similar evaluations are found in [35] and [36]. Based on keyword extraction, the accuracy of document summarization for the considered aspects is found as follows [35]:

For Literature: 70%
For Entertainment: 80%
For Sports: 76%

Based on sentence ranking [36] for categorizing the text, the compression ratios are as follows:

At 30% Compression ratio: 80%
At 40% Compression ratio: 83.33%

8. CONCLUSION

Abstractive summarization has not been considered in full extent in Kannada. There are many issues to be addressed in this context. Knowledge of the different categories is a prerequisite for aspect selection. The handwritten IE rule repository must be exhaustive to handle the various syntactic properties of the language. Standardization of words and their spellings is a huge part of pre-processing. The usage of templates ensures that a clean and to the point summary is generated which is syntactically accurate. But monotony in the generated summaries is a possible side effect of using template based sentence generation.

9. FUTURE WORKS

The system can be expanded to include more domains. A possible variation of the system can be to produce the summary output in a more universal language like English instead of the result being in the same language as the original document. The results produced by the system can also be in the form of a speech output thus reducing the time needed to absorb the key facts in a document. A machine learning approach can be considered that would truly automate the entire system starting from the POS tagging to the sentence generation stage. A sentence generation system needs to be developed based on the grammar rules of Indian regional languages.

REFERENCES

1. Mari-Sanna Paukkeri, & Timo Honkela,'Likey, (2010). Unsupervised Language – independent Key phrase Extraction, Proceedings of *the 5th International Workshop on Semantic Evaluation*, (pp 162– 165), ACL 2010, Uppsala, Sweden.
2. The Karnataka Official Language Act, *Official website of Department of Parliamentary Affairs and Legislation*, Government of Karnataka. Retrieved 2007.
3. *Declaration of Telugu and Kannada as classical languages*, Press Information Bureau. Ministry of Culture, Government of India. Retrieved 2013.
4. Inderjeet Mani, (2001). Recent Developments in Text Summarization, CIKM-2001, Atlanta, Georgia, USA.
5. Te-Min Chang, & Wen-Feng Hsiao (2008). A hybrid approach to automatic text summarization, 8th IEEE International Conference on Computer and Information Technology 2008. pp 65 – 70.
6. Pierre – Etienne Genest, & Guy Lapalme, (2012). Fully Abstractive Approach to Guided Summarization, Proceedings of the *50th Annual Meeting of the Association for Computational Linguistics*, pp 354 – 358 , Jeju Republic of Korea, Association for Computational Linguistics.
7. Jackie CK Cheung, (2008), Comparing Abstractive and Extractive Summarization of Evaluative Text: Controversiality and Content Selection, B.Sc. (Hons.) *Thesis in the Department of Computer Science of the Faculty of Science*, University of British Columbia.
8. Yutaka Matsuo Mitsuru Ishizuka, (2004). Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information, *International Journal on Artificial Intelligence Tools*, Volume 13.
9. Dragomir Radev, Timothy Allison, Sasha Blair- Goldensohn, John Blitzer, Arda C¸ elebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, & Zhang Zhu., (2004). MEAD - a platform for multi document multilingual text summarization, *In Procs of LREC* Lisbon, Portugal.
10. Kaikhah, & Khosrow, (2004) Text Summarization Using Neural Networks, *Faculty Publications-Computer Science*. From https://digital.library.txstate.edu/handle/10877/3819.
11. Nadira Begum, Mohamed Abdel Fattah, & Fuji Ren, (2009). Automatic text summarization using support vector machine, *International Journal of Innovative Computing*, Volume 5, pp: 1987-1996.
12. Jayashree R, Srikanta Murthy K & Sunny K, (2011). Keyword Extraction Based Summarization of Categorized Kannada Text Documents, *International Journal on Soft Computing (IJSC)* Vol.2, No.4.
13. Trevor Cohn & Mirella Lapata, (2009). Sentence compression as tree transduction, *Journal of artificial intelligence*, 34(1):637–674.
14. Regina Barzilay, Kathleen R, & McKeown, (2005). Sentence fusion for multi document news summarization, *Computational Linguistics* 31(3):297–328.
15. Pierre-Etienne Genest & Guy Lapalme, (2011). Framework for Abstractive Summarization using Text-to-Text Generation, In Proceedings of *the Workshop on Monolingual Text-To-Text Generation*, pp. 64–73, Portland, Oregon, USA, Association for Computational Linguistics.
16. Rasim Alguliev & Ramiz Alguliev, (2009). Evolutionary Algorithm for Extractive Text Summarization, *Intelligent Information Management*, doi:10.4236/iim.2009.12019 pp. 128-138, Institute of Information Technology, Azerbaijan National Academy of Sciences, Baku, Azerbaijan.
17. Atif Khan & Naomie Salim, (2014). A review on abstractive summarization, *Journal of Theoretical and Applied Information Technology*, Vol. 59 No.1, ISSN: 1992-8645E-ISSN: 1817-3195.
18. Jagadish. S. Kallimani, Srinivasa. K. G, & Eswara Reddy B., (2011). Information extraction by an abstractive text summarization for an Indian regional language, IEEE 7th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), pp. 319-322. ISBN: 978-1-61284-729-0. INSPEC Accession Number: 12495964, Tokushima, Japan. DOI: 10.1109/NLPKE.2011.6138217.
19. Antony. P. J., & Soman. K. P., (2010). Kernel based part of speech tagger for Kannada, *International conference on Machine Learning and Cybernetics (ICMLC)*, vol.4, no., pp. 2139-2144.
20. Shambhavi. B. R. & Ramakanth Kumar. P., (2012). Kannada Part-Of-Speech Tagging with Probabilistic Classifiers, *International Journal of Computer Applications*. 48(17). pp. 26-30, Published by Foundation of Computer Science, New York, USA.
21. R. Grishman, Beth Sundheim., (1996). Message Understanding Conference-6: A Brief History in the procs of *the 16th International Conference on Computational Linguistics (COLING)*, pp: 466-471, Centre for Sprogteknologi, Copenhagen, Denmark.
22. H. Wallah, (2004). Conditional Random Fields: An Introduction, *University of Pennsylvania CIS Technical Report* MS-CIS-04-21.
23. J. Lafferty., A. McCallum., & F. Pereira, (2001). Conditional random fields: probabilistic models for segmenting, *International Conference on Machine Learning*.
24. Maynard. D., Tablan. V., Ursu. C., Cunningham. H. & Wilks. Y,, (2001). Named entity recognition from diverse text types, *Recent Advances in Natural Language Processing*. pp. 257-274.
25. Minkov Einat., Richard. C. Wang., & William. W. Cohen., (2005). Extracting personal names from email: applying named entity recognition to informal text, Proceedings of *the conference on Human Language Technology and Empirical*

Methods in Natural Language Processing. Association for Computational Linguistics.

26. A. Gatt., & E. Reiter., SimpleNLG (2009). A realisation engine for practical applications, *Proceedings of ENLG.*
27. Kees van Deemter, Emiel Krahmer, & Marie¨t Theune, (2005). Real versus Template-Based Natural Language Generation: A False Opposition*? Journal of Computational Linguistics*, Volume 31, Issue 1, pp. 15-24.
28. Jayashree. R, Srikanta. M. K., Anami. B. S., (2012). Categorized Text Document Summarization in the Kannada Language by sentence ranking, *12th International Conference on Intelligent Systems Design and Applications (ISDA)*, vol., no., pp. 776-781.
29. S. Teufel, (2001). Task-based evaluation of summary quality: describing relationships between scientific papers, in Proceedings of *the NAACL Workshop on Automatic Summarization*, pp. 12–21.
30. A. Nenkova., & K. McKeown., (2011). Automatic Summarization: Foundations and Trends in Information Retrieval, Vol. 5, Nos. 2–3, pp. 103–233.
31. Piwek Paul, (2003). A flexible pragmatics-driven language generator for animated agents, Proceedings of the *tenth conference on European chapter of the Association for Computational Linguistics*, Volume 2, Association for Computational Linguistics.
32. Siva Reddy and Serge Sharoff, (2011). 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.
33. Sujan Kumar Saha, Sudeshna Sarkar, & Pabitra Mitra, (2009). Gazetteer Preparation for Named Entity Recognition in Indian Languages, *The 6th Workshop on Asian Language resources.*
34. Jie Xu, Liyuan Xing, Perkis A, & Yuming Jiang, (2011). On the Properties of Mean Opinion Scores for Quality of Experience Management, *2011 IEEE International Symposium on Multimedia (ISM)*, doi: 10.1109/ISM.2011.88.
35. R Jayashree, K M Srikanta & K Sunny, (2011). Document Summarization in Kannada using Keyword Extraction, *Proceedings of AIAA 2011*, CS & IT 03, pp. 121–127.
36. R. Jayashree, (2012). Categorized Text Document Summarization in the Kannada Language by Sentence Ranking, *Proceedings of 12th International Conference on Intelligent Systems Design and Applications (ISDA)*, pp. 776-781.

---

[1] The { } bracket contains WX notation equivalence for Kannada character.

[2] The [ ] bracket contains English language word corresponding to Kannada language word.