Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation, pages 60–65
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Abstractive Text Summarization for Sanskrit Prose: A Study of Methods and Approaches

Shagun Sinha, Girish Nath Jha
School of Sanskrit and Indic Studies
Jawaharlal Nehru University, New Delhi
{shagunsinha5, girishjha}@gmail.com

Abstract

The authors present a work-in-progress in the field of Abstractive Text Summarization (ATS) for Sanskrit Prose – a first attempt at ATS for Sanskrit (SATS). We will evaluate recent approaches and methods used for ATS and argue for the ones to be adopted for Sanskrit prose considering the unique properties of the language. There are three goals of SATS - to make manuscript summaries, to enrich the semantic processing of Sanskrit, and to improve the information retrieval systems in the language. While Extractive Text Summarization (ETS) is an important method, the summaries it generates are not always coherent. For qualitative coherent summaries, ATS is considered a better option by scholars. This paper reviews various ATS/ETS approaches for Sanskrit and other Indian Languages done till date. In the preliminary overview, authors conclude that of the two available approaches - structure-based and semantics-based - the latter would be viable owing to the rich morphology of Sanskrit. Moreover, a graph-based method may also be suitable. The second suggested method is the supervised-learning method. The authors also suggest attempting cross-lingual summarization as an extension to this work in future.

Keywords: Abstractive Text Summarization, Sanskrit Prose, Computational Linguistics

1. Introduction

Text Summarization (TS) is a core area of study under Computational Linguistics (CL) and Natural Language Processing (NLP) for generation of coherent text summaries. One of the earliest works was by Luhn (1958) from IBM where he proposed to create summaries of the abstracts of scientific papers. TS has also been developed for a number of Indian Languages (ILs). Extractive text summarization (ETS) and abstractive text summarization (ATS) are two primary approaches that focus on summarizing IL internet content, newspaper articles, research papers, official documents etc (Sankar et al., 2011; Embar et al., 2013; Talukder et al., 2019; Gupta & Lehal, 2011; so on). Sanskrit is studied in various forms today mostly as a compositional language preserving several million texts of great intellectual value. The issues of text availability, readability and the need to access the knowledge in it have presented a huge requirement for ATS and related research for Sanskrit. The capacity of Sanskrit to infinitely condense an expression with recurrent usage of concatenating techniques like euphonic combinations (sandhi), compounding (samasa), scrambling, verb elision for prosody etc make it difficult to arrive at the structural or collocational meaning of the expression. When creating summaries, it is important that the semantics is processed well. Doing a good ATS for Sanskrit thus becomes extremely challenging. Summarization can be categorized differently on different bases: Single versus multi-document (based on the number of documents (Jones, 1999), textual versus multimedia (based on the mode of document), extractive versus abstractive (based on the mode of the output (Afantenos et al, 2005; Moawad & Aref, 2012). This paper is the description of an ongoing work on Sanskrit ATS (SATS) by the authors. The main contribution of this paper lies in its surveying the existing approaches to TS done for Sanskrit till date and to look at some challenges in processing Sanskrit for ATS. The paper proposes a semantic approach for any deeper processing of the texts in the language. The authors focus on single document summarization only because a multi-document ATS may be more complex due to various factors like semantic relatedness, diversity of subject matter, size etc.

2. Motivation for Sanskrit ATS

The origin and development of TS was inspired by the need to turn long English scientific texts into shorter ones (Luhn, 1958). Currently, most ideas around TS techniques under Natural Language Processing are based on the growth of the internet and the need to condense information therein (Sunitha et. al., 2016). In this backdrop, it is important to make one observation. While Sanskrit prose content on the net needs to be summarized as well, there are two key objectives of SATS which are different from those of TS in any other language of the present day:

- A large body of scientific literature is available in Sanskrit and a lot of it is in the manuscript (MS) form. The study of an MS is a far more complex and tedious process which involves editing and re-editing a historical document till the authentic content is achieved.

- SATS will require semantic analysis. This could pave the way for better semantic processing of Sanskrit. Since ATS works on the principle of 'key essence' of the text rather than extracting the suitable sentences, it could help enhance algo-

rithms for processing the relative meaning of the words.

3. Literature Survey

Sanskrit TS so far has explored the extractive aspect only. Barve et.al. (2015) use three TS approaches to obtain text summary for Sanskrit based on a query given by the user - Average Term Frequency-Inverse Sentence Frequency (tf-isf), Vector Space Model (VSM), and Graph-Based Approach. They concluded that the VSM produced the best summary with 80% accuracy. ETS is a good approach for prose that has a high frequency of the query-word, as is seen in Barve et. al (2015). However, not all prose may yield such results. In most cases, the keyword is not always repeated but is indicated through pronouns. While query-directed extraction can be highly successful in the former, it may not be so for the latter. Besides, the ETS also faces the incoherence disadvantage as mentioned by Mishra & Gayen (2018). Abstractive approach, on the other hand, is more likely to resolve this. It 'abstracts' the essence from the text to be summarized. This leads to complexity in language processing but once successful, can result in enhanced summary quality with natural text generation. Scholars suggest that non-extractive methods generate better summaries because they reduce the information loss (Mishra & Gayen, 2018). ATS has also been found better than ETS in other work (Giuseppe & Jackie, 2008).

3.1. Major ATS approaches for Indian Languages:

Scholars have different bases for organizing the types of TS. Most of them can come under one or more of these categories:

1. Structure vs Semantic approach (Sunitha C et al., 2016),
2. Machine Learning (ML) based methods (Anh & Trang, 2019; Talukder et al., 2019) , and
3. Corpora based approach (Hasler et al., 2003)

3.1.1.

Sunitha C et. al. (2016) present a survey of the current techniques in ATS for ILs. Key approaches to ATS in ILs can be divided into two categories: Structure-based and Semantics based. Some notable works in ILs include Rich Semantic Graph approach for Hindi (Subramaniam & Dalal, 2015), Malayalam (Kabeer & Idicula, 2014), ATS through an extractive approach for Kannada (Kallimani et. el, 2014).

Structure-based approaches require the source text sentences to be collected in a predefined structure (Sunitha et al, 2016). The types of structures mentioned are Tree-based, Ontology-based, Lead and Phrase structure based, Rule based and Template-based. Each of these methods aims to collect the sentences from the source text and then generate a summary later.

In the Semantics based approach, there are three phases that lead to the summary- document input, semantic review and representation and then finally summary based on this semantic representation through Natural Language Generation (Sunitha et al., 2016). Multimodal semantic, Information Item-based and Semantic Graph (Moawad & Aref, 2012) are the methods which focus primarily on the semantic representation of the source text. It is important to note that abstraction will need semantic representation at some stage. and that ATS requires two major components always - meaning extraction and summary generation in natural language.

A closer look reveals that the ILs popularly use : the graph-, the POS-NER-, and textual position-based methods.

Of the given types, one common method is the ontology based method. Ontology refers to the 'theory of existence' or a list of all the things that exist (Russell & Norvig, 2019). A number of such summarization tools have been developed for a field-specific summarization. For example, Texminer is a tool that summarizes papers of Port and Coastal Engineering (Hipola et al, 2014); or it may be related to a particular scientific field (Luhn, 1958). We find it noteworthy that ontology is important in areas where a finite set of vocabulary pertaining to the field can be enlisted.

However, in extraction techniques in NLP, a method of ontology extraction does exist (Russell & Norvig, 2019). This may be a possible approach to get some ontology out of a general document, but its reliability for summarization purposes may have to be tested.

This brings us to the next possible approach to Indian languages text summarization which is graph-based summarization. Graphs are created out of the text document with its words as vertices and the links between them as edges (Subramaniam & Dalal, 2015). This method can be used for languages with easy tokenization availability. An additional use of WordNet is also required here.

Advanced work in graph-based methods includes 'Reduced Semantic Graph' (RSG) methods where an even more simplified version of a text's graph is generated using ontology for word-sense instantiation, concept validation and sentence-ranking (Moawad & Aref, 2012). RSG methods have been deployed for Hindi (Subramaniam & Dalal, 2015) and Malayalam (Kabeer & Idicula, 2014). The results for Hindi are reported to be up to the mark (Subramaniam & Dalal, 2015).

Due to the rich morphology of Sanskrit, a standard word-order may not be followed even in current prose. Semantic representation thus becomes an essential element. This indicates that perhaps semantic approach would yield better results.

3.1.2. Machine Learning Approaches:

One other way of classifying the TS types is the ML based approach: supervised and unsupervised methods (Majid & Fizi-Derakashi, 2015). Supervised methods require texts with their labeled summaries for training.

Unsupervised methods include graph-based, VSM, text-based. Graph-based method can be grouped with the semantic graph approach mentioned earlier. It creates a graph with concepts as vertices and the relation between them as edges (Majod, & Fizi-Derakhshi, 2015).
VSM technique creates vectors of the units of text and then the most important units are extracted with the help of a semantic analysis technique (Maji & Fizi-Derakhshi, 2015).
A neural-network based application of the Memansa principle is used by Sakhare and Kumar (2016). Although they use it for English through neural nets, the approach for information extraction is taken from Mimamsa which makes it relevant to our discussion.
A pointer-generator method based on pre-trained word-embedding for ATS has been performed for English by Anh & Trang (2019). The application for Sanskrit will need to be tested though they had the prepared CNN/Dailymail dataset for training already.
Another effort in IL ATS has been by Talukder et al. (2019) where the model used is sequence to sequence RNN. They report the loss of training error to 0.008.
The text-based method is classified as the third method. This is the corpus-based method deployed by others (Hasler et al, 2003; Edmundson, 1969) discussed in the next section.
Apart from graph-based methods, POS-NER based methods have also been deployed. Embar et al (2013) presents sArAmsha, an abstractive summarizer for Kannada. According to them, tools like POS tagging and NER implementation are used in the initial processing of documents and then an abstraction scheme is applied. This may also be classified under the corpus based approach.

3.1.3. Corpus based approach:

Under this, Corpus is annotated with relevant annotation schemes like POS, NER, discourse annotation tools like the Rhetorical-Structure Theory (Mann & Thompson, 1988; Jones, 1999; Zahri et al., 2015) etc, which helps in extracting meaning at a later stage.
Corpus type has also been used as an important basis for developing TS (Hasler et al., 2003). Annotation of corpora to indicate meaningful units in a text is a viable method. The works suggest that semantic abstraction becomes easier with this annotated corpora. However, Oya T. et al. (2014) use template-based abstractive summarization which they report has reduced dependence on annotated corpora.

3.1.4.

At this point, it is important to mention the extraction-based abstraction approach to TS one of which is the Information Extraction(IE) ATS (Kallimani et al, 2011). IE techniques are deployed in the initial stages in order to identify the important word units in the document. Abstraction is done from these extracted units (Kallimani et al, 2011; Afantenos et al., 2005).
Edmundson(1969) used a proper corpus divided into training and testing for summarization and evaluation. The method used is feature based only and he suggested that it was important to consider syntactic and semantic features in summarization. It may be noted that the 'abstracting' referred to in his article is focused on generating abstracts of articles based on extracted sentences. He terms this process as 'abstracting' (Edmundson & Wylls, 1961), though it is different from abstraction as we know it today.
Other than ATS, some prominent works in ETS for Indian Languages have been covered by Dhanya & Jathavedan (2013). The latter includes the thematic and positional score based method for Bengali (Sarkar, 2012); statistical features like cue phrase, title keyword, and similar features based extraction method for Punjabi (Gupta & Lehal, 2011); the graph- based text ranking method for Tamil (Sankar et al, 2011) performs extractive summary without any annotated corpora or supervised learning method.
Patel et al. (2007) and D'Silva & Sharma (2019) look at multilingual translation problems with language independent TS being one option (Patel et al., 2007). There are two reasons why it may not be useful to us. First, their approach is statistical and not semantic. It has been suggested by Edmundson (1969) that syntactic and semantic factors as well as context of a text (Jones, 1999) in TS be considered for better quality. We too believe that semantic representation is important for ATS. Two, their approach is mostly extractive. The other option, that of cross lingual TS using Machine Translation (MT) (D'Silva & Sharma, 2019) is a good option to be explored.

3.1.5.

A key point to be observed in these and general text summarization tools is the type and source of data. There are two primary domains of data on which most tools are based: Scientific articles and newspaper articles. Tools for the summary of these two types of texts are usually developed more. While extractive is a dominant approach for these domains, abstractive has also a good presence.
However, to begin a process in Sanskrit ATS, we have focused our study on contemporary prose consisting of mainly newspaper articles and Sanskrit blogs.
Observations regarding methods:

1. Scholars use TS methods in a mixed manner. For e.g., a semantic graph may require ontology deployment for better semantic representation (Moawad & Aref, 2012); abstractive summarizer may first extract relevant information before applying abstraction (Kalimanni et al, 2011).

2. Supervised methods will need label summaries along with the texts. Thus, newspaper articles with their headlines are usually taken as the standard training corpus where the headline serves as the summary of the respective text. This is a feasible approach for a beginner-level work.

4. Sanskrit ATS

Some features of the Sanskrit writings and their challenges can be stated as following:

- Sanskrit prose is strictly based on the principles of grammar which inspires its word-formation and usage. Owing to the Paninian model of Grammar, the language is rich in morphology. The principle of economy and precision have been important for Sanskrit prose(Kiparsky, 1991). As a result, while the prose in Sanskrit in general is appreciated for its economy, it becomes difficult for any man/machine processing, and more so for the ATS.

- **Compounds and Sandhis**: Sanskrit prose is constituted on the samhita (continuous text) principle thereby using Compounds and Sandhis (euphonic combination) heavily. For instance, multiple words combined after removing their inflections is an example of a compound. Space does not act as a delimiter largely here. This along with potentially recursive sandhi and complex morphology make preprocessing a critical task for Sanskrit texts.

- **Word Significance**: Most Sanskrit literary works, especially poetry, tend to be indirect in their intended meanings - abhidha(literal), lakshana(metaphor), vyanjana(euphemism). Poetry usually expresses meanings more than one but the same can go for most prose creations in literature also. The availability of lexical resources like the Amarakosha bear testimony to this fact, so does the long tradition of language analysis including the philosophy of Mimamsa (interpretation) and Nyaya (logic).

- **Diversity of verb usage**: While lakaras (tense) are used to denote time, some suffixes are also used to indicate past and present tense. Thus, for the same verb, different forms of it can be used to suggest the same meaning. For each such usage, meaning will have to be considered well before generating a summary of any type.

5. Preliminary Study

To perform a preliminary data study, a total of 1310 sentences have been extracted from online sources and stored as data files. Current prose like the news articles from the All India Radio, DD News and other sources have been considered at this stage. The following may be observed about the data:

1. Sentences are usually short, with not more than 7 words per sentence on an average.

2. Owing to the fact that most digital sources in Sanskrit found so far exist as a way to teach prospective learners, there is no variety in content found there.

3. News articles offer a good standard of sentences in Sanskrit while at the same time reducing the complexity of verbs. There are a few standard usages which ensure ease of meaning comprehension.

The short length of sentences indicates that with some basic preprocessing only, a TS method may be applied on the text. After going through the preliminary data, this has led us to conclude that we may start our work with focus on two approaches: first, a graph-based method. Owing to short sentences in the current prose, generating a graph and the prospective relations among words may be quicker and efficient.
Second, supervised method where news articles and their headlines are taken as corpora for training. This would be on the lines of the ATS developed on English and other languages using the CNN/Dailymail dataset (Mishra & Gayen, 2018).
Preprocessing of the text is a necessary stage in the approach (Barve et. al, 2015). This would ensure creation of words for ease of processing the text further. Contemporary simple prose that contains direct meanings instead of oblique ones should be used like Barve et al (2015) use Sanskrit Wikipedia articles to test their approaches (VSM, Graph and tf-isf).
A work on these two methods will suggest further course of action. Annotation may be required if the results so indicate.

6. Conclusion

This paper presents a preliminary attempt to develop a Sanskrit abstractive text summarizer for current prose. It surveyed the top abstractive summarization approaches to Indian languages, in general, with a view to zeroing in on one approach for the current work on Sanskrit ATS. Since there has not been any attempt at Sanskrit ATS so far, a beginning is being made for current Sanskrit prose mostly news articles. While summarization would not suit literary poetry, we could utilize dependency parsers to build semantic graphs for any verse in scientific texts. Prose in these texts could be further summarized if this work is advanced further from current prose to other prose styles. After surveying the available literature for ATS in ILs the authors propose that semantic approach would be better suited for the inherent complexities that Sanskrit is known for. Owing to rich morphology of the language, pre-defined structures may not result in a coherent or usable summary. Thus, a semantic approach would assist in arriving at a better analyzed summary. In the semantic approach, a graph-based method shall be a good start. Secondly, a supervised method for the available prose from the news article-headline combine may be emulated for Sanskrit too.
The possibility of annotation should be considered after this, if required.
The language of the output summary is one dimension of SATS which is out of the scope of this paper. For any other language, the abstracted summary is produced in the same language as the text. However, it

could be explored if the abstractions of Sanskrit prose could be carried out in both Sanskrit as well as Hindi or English with the help of an existing Machine Translation.

7. References

Afantenos, S., Karkaletsis, V. & Stamatopoulos, P. (2005). Summarization from Medical Documents: A Survey. Artificial Intelligence in Medicine. 33. 157-177. 10.1016/j.artmed.2004.07.017.

Anh, D. T., & Trang, N. T. T. (2019, December). Abstractive Text Summarization Using Pointer-Generator Networks With Pre-trained Word Embedding. In *Proceedings of the Tenth International Symposium on Information and Communication Technology* (pp. 473-478).

Barve, S, Desai, S. & Sardinha, R. (2015). "Query-Based Extractive Text Summarization for Sanskrit". In: *Proceedings of the Fourth International Conference on Frontiers in Intelligent Computing: Theory and Applications(FICTA)*. Springer. Digital Object I: 10.1007/978-81-322-2695-6_47

C. Sunitha, A., Jaya, & Ganesh, A. (2016). "A Study on Abstractive Summarization Techniques in Indian Languages". In: *Proceedings of the Fourth International Conference on Recent Trends in Computer Science and Engineering.* Procedia Computer Science. 87(2016). pp 25-31. Elsevier: DOI: 10.1016/j.procs.2016.05.121

D'Silva, J. & Sharma, U (2019). Automatic Text Summarization of Indian Languages: A Multilingual Problem. *Journal of Theoretical and Applied Information Technology. 97(11).*

Embar, V., Deshpande, S., Vaishnavi, A.K. & Jain, V. & Kallimani, J. (2013). sArAmsha - A Kannada abstractive summarizer. In: *Proceedings of the 2013 International Conference on Advances in Computing, Communications and Informatics*, *ICACCI* 2013. 540-544. 10.1109/ICACCI.2013.6637229.

Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM (JACM)*, *16*(2), 264-285.

Edmundson, H. P., & Wyllys, R. E. (1961). Automatic abstracting and indexing—survey and recommendations. *Communications of the ACM*, *4*(5), 226-234.

Gupta, V., & Lehal, G.S. (2011). Features Selection and Weight learning for Punjabi Text Summarization. *International Journal of Engineering Trends and Technology. 2(2).*

Giuseppe C & Jackie C. K. (2008), Extractive vs. NLG-based Abstractive Summarization of Evaluative Text: The Effect of Corpus Controversiality, *Proceedings of the Fifth International Natural Language Generation Conference, ACL,* https://www.aclweb.org/anthology/W08-1106

Hasler, L., Orasan, C., & Mitkov, R. (2003). Building better corpora for summarization. In *Proceedings of Corpus Linguistics* (pp. 309-319).

Hipola, P., Senso, J.A., Mederos-Leiva, A. & Dominguez-Velasco, S. (2014). Ontology-based text summarization. The case of Texminer. *Library HiTech. 32(2)*. pp 229-248. Emerald. DOI: 10.1108/LHT-01-2014-0005.

Jones, K. S. (1999). Automatic summarizing: factors and directions. In Mani & Maybury (eds.) *Advances in automatic text summarization* (No. 1, pp. 1-12). Cambridge, Mass, USA: MIT press.

Kallimani, J. S., & Srinivasa, K. G. (2011). Information extraction by an abstractive text summarization for an Indian regional language. In *2011 7th International Conference on Natural Language Processing and Knowledge Engineering* (pp. 319-322). IEEE.

Kabeer, R. & Idicula, S. M.(2014). "Text summarization for Malayalam documents - An experience" In: *Proceedings of the International Conference on Data Science & Engineering* (ICDSE), Kochi, pp. 145-150.

Kiparsky, P. (1991). Economy and the Construction of Sivasutras. PDF.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, *2*(2), 159-165..

Mani, I. & Maybury, M. T. (1999). *Advances in Automatic Summarization.* MIT Press.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, *8*(3). pp. 243-281.

Moawad, I F & Aref. M. *(*2012). "Semantic Graph Reduction Approach for Abstractive Text Summarization". In: *ICCES.* p 132-138. DOI: 10.1109/ICCES.2012.6408498

Mishra, R. and Gayen, T. (2018). "Automatic Lossless Summarization of News Articles with Abstract Meaning Representation." In: *Proceedings of the 3rd International Conference Computer Science and Computational Engineering.* Procedia Computer Science. PDF.

Oya, T., Mehdad, Y., Carenini, G., & Ng, R. (2014). A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*: pp. 45-53.

Patel, A., Siddiqui, T., & Tiwary, U. S. (2007). A language independent approach to multilingual text summarization. *Large scale semantic access to content (text, image, video, and sound)*, 123-132.

P.M, Dhanya & Jathavedan M. (2013). “Comparative Study of Text Summarization in Indian Languages.” In: *International Journal of Computer Applications.* 75(6) : pp 17-21.

Ramezani, M. & Feizi-Derakhshi, Md. R. (2015). Ontology-Based Automatic Text Summarization using FarsNet. *Advances in Computer Science: an International Journal.* 4(2) no.14.

Russell, S J. & Norvig, P. (2019). *Artificial Intelligence: A Modern Approach.* Pearson.

Sankar, K., R, Vijay Sundar Kumar, Devi, S.L. (2011). Text Extraction for an Agglutinative Language. *Language in India. 11(5). Special Vol: Problem of Parsing in Indian languages.*

Sakhare, D.Y. and Kumar R (2016). Syntactical Knowledge and Sanskrit Memansa Principle Based Approach for Text Summarization” In: *International Journal of Computer Science and Information Security (IJCSIS).* 14(4). pp. 270-275. ISSN: 1947-5500.

Sarkar, K. (2012). Bengali text summarization by sentence extraction. *arXiv preprint arXiv:1201.2240*.

Subramaniam, M. & Dalal V. (2015). “Test Model for Rich Semantic Graph Representation for Hindi Text using Abstractive Method” In: *International Research Journal of Engineering and Technology.* 2(2). pp 113-116. e-ISSN:2395-0056

Talukder, M. A. I., Abujar S., Masum, A. K. M., Faisal, F. & Hossain, S. A. (2019). "Bengali abstractive text summarization using sequence to sequence RNNs," *10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. pp. 1-5.

Zahri, N.A.H., Fukumoto, F., Suguru, M., & Lynn, O.B. (2015). Applications of Rhetorical Relations Between Sentences to Cluster-Based Text Summarization. in Nagamalai et al. (eds.) *CCSEA, DKMP, AIFU,* SEA-2015. pp. 73-92. 10.5121/csit.2015.50207.