International Journal
of
Linguistics & Computing Research

Journal Homepage: http://cphfs.in/research.php

Machine Learning to Traditional War-form: Extracting Semantic Knowledge from "Sarala Mahabharata"

Rudranarayan Mohapatra
P.G. Department of Odia,
Utkal University, Vanivihar, Bhubaneswar,
E-mail: rudrautkal@gmail.com

***Abstract:** Natural language processes and machine learning technologies have acquired considerable importance, especially in disciplines of classical and heritage documents. The primary information in classical heritage documents is generally available in free text form. Therefore, the present paper is an initiative to carry out the cultural and heritage text materials to a relational database form that could be capable of browsing big online knowledge repositories. With this, the system could be educating itself as on demand specially to traditional War-form and used for producing semantic metadata ready to be integrated with information coming from Odia heritage documents, to establish an advanced machine learning scenario. The ultimate goal of this paper is to convert "Sarala Mahabharata" into a digital humanities resource.*

Key Words: H*eritage Preservation, Digital Humanities, Sarala Mahabharata, manuscripts*

I. INTRODUCTION

The dream of having machines capable of reading a text automatically by understanding its in-depth semantics and pragmatics, is as old as computer science. Its realization would cater our need of acquiring information formerly inaccessible or hard to access. Again, this would also benefit the knowledge bases of many disciplines in higher form. However, today, there are powerful tools capable of reading a text and are now working with most major world languages to perform complex operation both in linguistics as well as in technology. Out of the tasks, notably language detection, parts of speech (POS) recognition and tagging, grasping grammatical, syntactic and logical relationships are being used for Natural Language Processing (NLP). To a certain extent, they can also speculate, in very general terms, on the meaning of each single word (e.g., whether a noun refers to a person, a place or an event). Named entities resolution (NER) and disambiguation techniques are the ultimate borderline of our NLP research. Beyond which we have to think over the actual possibility of making the machines aware of the semantic knowledge of a text. Recently, the "Venice Time Machine project" (vtm.epfl.ch), aimed at (re)writing the history of the city by means of "stories automatically extracted from of ancient manuscripts", has greatly contributed to feeding this interest and rendering it topical.

Looking to the heritage prospective of Odisha and Odia language, the great epic, 'Sarala Mahabharata' stands out as an independent, autonomous piece of art and classic on its own merit. Though, Krishna-Dwaipayana Vyasa's original Sanskrit Mahabharata remains the raison ancestry of Sarala's Mahabharata in Oriya, it is not at all a translation of the former or even written in the shadow of it. A closer examination of Sarala's epic would expose its five broad patterns: Religious, Socio-Cultural, and Ethical intellectual-philosophical, Structural and linguistic knowledge that is an asset for present and for future.

II. DIGITAL HUMANITIES & HERITAGE PRESERVATION

In the era of big data analysis, digital humanities face the ongoing challenge of formulating a long-term and complete strategy for creating and managing interoperable and accessible database to support its research aims. Though, the semantic analysis and formal ontological hierarchy someway properly understood and provide a powerful potential solution to this problem. A long-term research programme for representing big data is key areas of research in the digital humanities, aiming to support the sustainable development of this interoperable and accessible datasets. According to Berners-Lee, semantic data refers to data which is machine processable and human readable. Formal ontologies provide explicit and disciplined means of producing such data, to ensure its wide compatibility and clear interpretability. Therefore, the time has come to preserve the Ancient manuscripts including the manuscripts of 'Sarala Mahabharata', the one of the primary sources of Odisha cultural heritage.

III. INDIAN MANUSCRIPTS ARE THE RICHEST COLLECTION

As per the study, total number of manuscripts in India 5 million. Out of them, Indian manuscripts available in European countries are 60,000. Indian manuscripts in South Asia and Asian countries are 1, 50,000. Number of manuscripts recorded in catalogues 1 million. However, Percentage of manuscripts in language wise are: Sanskrit-67%, Other Indian languages-25% and Arabic/ Persian/ Tibetan-8%

Out of the above, in Odisha the manuscripts preserved with us are of about 10,000 plus and it scattered as follows:

- Parija Library Utkal University, VaniVihar - 6000
- P.G Dept. of History, Brahmapur University, BhanjaVihar – 989
- PuribadaodiaMatha – 450
- P.G Dept. of OdiaSambalpur University -315
- KedarnathGabesanaPratistana, Bhubaneswar – 265
- Raghunandana Library. Puri – 250
- Utkal Sanskruti Viswavidyala – 95
- P.G Dept. of History, Sambalpur Universiy- 1480

The above collections have subjects covering almost all the sub-branches of Sanskrit and Classic Odia literature.

IV. IMPORTANCE OF DIGITIZATION

The digitization of our ancient manuscripts has many more importance as it is faster to Access and it can improve services. Easy to Archiving and protection of the originality of the object are also possible. It also reduces the handling and use of fragile or heavily used original material. The Resource Sharing and preservation is also going to be relatively easy.

V. IMPORTANCE OF DIGITAL PRESERVATION

- Preservation links the past with the future
- Knowledge is one of the few things that lasts
- Significant part of world's knowledge & heritage is in digital form
- E-resources can and do disappear.
- Protect from Loss: Example Location of NASA's original moon landing recordings is (currently) unknown. Same case to Our Odisha Rasagolla issue.
- Orphans: When ownership or other rights become uncertain, availability is threatened

VI. SOME INITIATIVES IN DIGITAL PRESERVATION (WORLD & INDIA)

- Digital Information Archiving System (DIAS) developed by IBM.
- 'Planets' is a Network services co-funded by European Union to addree core digital preservation challenges.
- kopal (Co-operative Development of a Long -Term Digital Information Archive) developed for long-term accessibility of digital documents. This system is jointly devised by IBM and the National Library of The Netherlands in The Hague.
- National Digital Preservation Programme is lunched by Department of Information Technology, Govt. of India for the purpose of long term digital preservation of historically significant cultural materials, heritage archives, citizen information etc.

With this, the Indira Gandhi National Centre for the Arts (IGNCA) has taken a number of remarkable steps in the area of art and culture comprising the fields of creative and critical literature, written and oral; the visual and performing arts. A unique feature of the Reprography unit of the "Kalanidhi (National Information System and a Data Bank of IGNCA)" is the reprographic compilation of unpublished manuscripts in Indian and foreign collections from private and public libraries. A pioneering attempt has been made to bring under one roof primary sources of the Indian tradition lying scattered, fragmented, inaccessible or worse, in danger of extinction. At present the library contains more than one million folios of unpublished Sanskrit, Pali, Persian and Arabic manuscripts.

VII. THE LIFECYCLE OF DIGITAL PRESERVATION

The lifecycle of digital preservation is shown at Fig.No.1 in the Appendix.

VIII. EXEMPLARY TEXT FROM SARALA MAHABHARATA FOR WAR-FORM DATA ANALYSIS

To analyze the war-form data of Mahabharata Yudha, we extracted the portion of text from "Kaurava Pandabankara Yudhyasajja" and "Kuru Pandaba sainyasaha" also from "Bhismaparva" of Sarala Mahabharata. The text is as at Table No.1.

With this the, analysis portion of text contains 67 Characters (Name of persons including deities) in the mode of communication. Therefore, the whole characters involvement to context of the whole text demands an ontological relationship chain for machine learning purpose for deep structure analysis.

IX. METHODOLOGY

The methodology we apply to analyze these war data is first by text normalization process. Then we plan to adopt the COMPRENO parser like process which automatically converts text into a forest of syntactic-semantic trees which comprise dependency links and constituency structure.

In text Normalization process, we first try to find out the proper Nouns where the single object replaced by multiple names. Exemplary Normalization process is at **Table No.2**

X. ANAPHORA RESOLUTION

Due to unavailability of proper Anaphora resolution system, here, we are trying to manually tag and replace the pronoun to cater the deep semantic representation of text and dependency relationship. The dependency relationship text replacement is at **Table No. 3.**

Here, the both children nodes - one (or more) with 'Agent' deep syntax slot and another with 'Addressee' slot are linked node with a semantic class 'Verbs_Of_Communication'. And, the verbal arguments differentiate on the basis of their syntactic position within the verbal semantic frame - Experiencer, Agent, Patient, Addressee, and Possessor.

The analysis is based on the universal semantic hierarchy—a complex WordNet-like ontological structure that stores meanings rather than words (Manicheva et al., 2012; Petrova, 2013). The resulting trees contain nodes with all sorts of linguistic information attached to them: semantic classes from the said hierarchy (e.g. 'Person_By_Firstname' or 'Verbs_Of_Addressing'), purely syntactic 'surface slots', syntactic-semantic 'deep slots' (e.g. Agent or Experiencer)'.

The final list of roles included Agent, Object (equivalent to Patient), Experiencer, Addressee, and Possessor demonstrates the standardized results of the semantic role distribution of Character 'ShriKrishna'

The distribution is at **Table No. 4**.

The above representation shows the continuation of facts as applicable in real context.

CONCLUSION

The above deep-structure analysis shows that there are certain dependencies between the apparent personal traits of a character and his or her positions within the deep structures. We hope that further research will help us gain more insights into the 'literary technique' of 'Sarala Mohabharata' and enables us to create a semantic mark-up of his works. Again, the semantic analysis if war-data can be possible by digitized and digital preservation of our manuscripts.

REFERENCE:

1. Saarala MahaBharata, VishmaParva, Publisher: Sarala Sahitya Sansada
2. Gaur, Dr. Ramesh C, 'Digitization and Digital Preservation of Indian Cultural HeritageIndian Heritage', HeritageMultimedia Digital Library Initiatives at IGNCA, New Delhi
3. Das, Dr. Satyabrat, 'Sarala Mahabharat: A Study', Orissa Review, June-July 2007
4. Roy, Pratap Chandra, 'The Mahabharata or Krishna-Dwaipayana Vyasa', (Translated into English prose from the Original Sanskrit Text', Vol. I, Adi Parva, This free e-book has been downloaded from www.holybooks.com:http://www.holybooks.com/mahabharata-all-volumes-in-12-pdf-files/
5. Sahoo. Jyotshna, 'Indian Manuscript Heritage and the Role of National Mission for Manuscripts', University of Nebraska - Lincoln DigitalCommons@University of Nebraska - Lincoln, http://digitalcommons.unl.edu/libphilprac
6. Luca Pezzati and Achille Felicetti (INO-CNR), 'DIGILAB: A New Infrastructure for Heritage Science', The European Research Infrastructure for Heritage Science, E-RIHS
7. Saptarshi Kolay, 'Cultural Heritage Preservation of Traditional Indian Art through Virtual New-media', "Conservation of Architectural Heritage, CAH" 23-27 November 2015, Luxor, ScienceDirect
8. [1]Jain, Anil Kumar &et.al. 'Rare handwritten manuscript collection in Indic Languages at Scindia Oriental Research Institute (SORI), (India)' IFLA WLIC 2013, Submitted on June 1, 2013, this work is made available under the terms of the Creative Commons Attribution 3.0 (Unported License: creativecommons.org/licenses/by/3.0/), (Central Library, Banaras Hindu University, Varanasi 221005, Uttar Pradesh) http://library.ifla.org/17/1/095-jain-en.pd

APPENDIX

Fig. No.1: Lifecycle of Digital Preservation

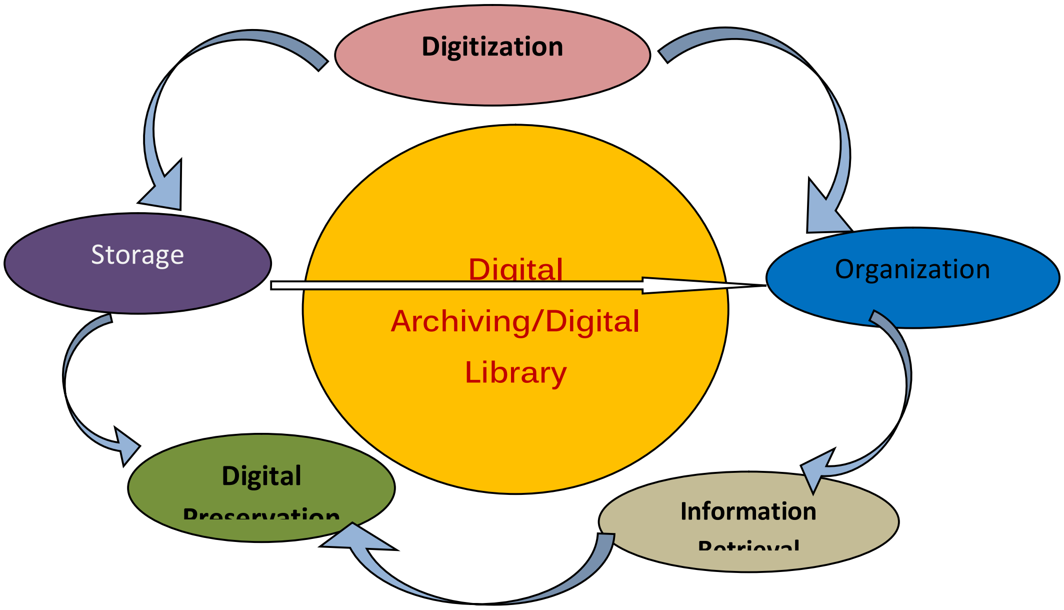

Table No.1

	Number	**Reference Text (From Odia)**	**Reference Text (In transliterate form)**
Ratha (Chariots)	8	ଅଷ୍ଟରଥ ଜାତ ହୋଇଲେ ପିତା ମହଂକର ଆହୁତି ସୂର୍ଯ୍ୟମଣ୍ଡଳ ରଥ ଚନ୍ଦ୍ରମଣ୍ଡଳ ରଥ ପୁଷ୍ୟେକ ରଥ ସଂଚକେତି (115) ଖେଚରୀ ରଥ ମନଦଣ୍ଡ ରଥ ନନ୍ଦିଘୋଷ ରଥ ଆବର ତାଳଧ୍ୱଜ ରଥ ଯେ ଅଷ୍ଟରଥ ସେ ଯାଗରୁ ଉପଗତ (116) ଯେ ଅଷ୍ଟରଥ ଯାଗ କୁଣ୍ଡରୁ ବାହାର ସାତରଥ ଥୋଇଲେ ନେଇ କୁବେର ଭଣ୍ଡାର (117)	Astaratha jaata hoile pitamahankara aahuti Suryamandala ratha chandramandala ratha pushyeka ratha sanchiketi (115) Khechari ratha manadanda rath nandighosa rath Aaabara taaladhawja ratha ye astharatha se jaagaru upagata (116) Ye astaratha jaaga kundaru bahara Sataratha thoile nei kubera bhandara (117)

Horse	4 in Nandighasa	ସେ ଚାରି ଅଶ୍ୱ ଯେ ଅକ୍ଷୟେ ଅବ୍ୟୟେ
		ତାହାନ୍ତ ଆଣି ନନ୍ଦିଘୋଷ ଯୋଚିଲେ ଦେବରାୟେ (132)
		ଶଙ୍ଖ ଗୋକ୍ଷୀର ଯେ ଶ୍ୱେତ କାମପାଳ
		ଯେ ଚାରି ଅଶ୍ୱେ ନେଇ ଯୋଚିଲେ ଶୁଭବେଳେ ଦେବୀକର ବାଳ (134)
			Se Chari Asva je Akshye Abaye;
			Tahanta Aani nandighosa jochile Debaraye (132)
			Sankha Gokshira je Shweta Kaamapala;
			Ye Chari Aswe Nei Jochile Subhabele Debikara baala (134)

Arms and weapons		ଦୁତୀୟ ପିନାକ ଅକ୍ଷୟେ ତ୍ରୋଣ ବେନି କେଶବ ଖଞ୍ଜିଲେ ଯେ ଶ୍ରୀକରେ ତାହା ଘେନି (59) ପାତାଳୁଂ ମନଭେଦି ଶର ଦିଲେକ ବହତି ଶ୍ରୀଭୁଜେ ଧଇଲେ ତାହା ବିଚିତ୍ରବୀର୍ଯ୍ୟର ନାତି (60)	Trutiya Pinaaka Akshye trona beni Keshaba khanjile je shrikare taha gheni (59) Paatalun Manabhedi shara dileka bahati Shribhuje dhaile taha bichitrabiryare naati (60)

Table No.2

Main Text	**Proper Noun replacement**
Sahadeva Bachane ***Prabhunka*** Laagila chinta Dhyaane sumarile deva digapaala debata (98) ***Naarayan****a* Sumarante Samaste hoile drushya Binaya Bhaba hoina se samasta milile ***SriKrushn***a Pasa (99) ***Shrihari*** Boile ye mahhabhaarata samara, Mote Agnya deichhanti Brahma Shrimukhare (100)	Sahadeva Bachane ***Prabhunka*** **(ShriKrushna)** Laagila chinta Dhyaane sumarile deva digapaala debata (98) ***Naarayan****a* **(ShriKrushna)** Sumarante Samaste hoile drushya Binaya Bhaba hoina se samasta milile ***SriKrushn***a Pasa (99) ***Shrihari*** **(ShriKrushna)** Boile ye mahhabhaarata samara, Mote Agnya deichhanti Brahma Shrimukhare (100)

Table No. 3

Main Text	**Proper Noun replacement**
Sahadeva Bachane ***Prabhunka*** (***SriKrushn***a) Laagila chinta Dhyaane sumarile deva digapaala debata (98) ***Naarayan****a* (***SriKrushn***a) Sumarante Samaste hoile drushya Binaya Bhaba hoina se samasta milile ***SriKrushn***aNka Pasa (99) ***Shrihari*** (***SriKrushn***a) Boile ye mahhabhaarata samara, Mote Agnya deichhanti Brahma Shrimukhare (100)	Sahadeva Bachane ***Prabhunka*** (***SriKrushn***a) Laagila chinta Dhyaane sumarile deva digapaala debata (98) ***Naarayan****a* (***SriKrushn***a) Sumarante Samaste (***Deva Digapaala***) hoile drushya Binaya Bhaba hoina se samasta (***Deva Digapaala***) milile ***SriKrushn***aNka Pasa (99) ***Shrihari*** (***SriKrushn***a) Boile ye mahhabhaarata samara, Mote (***SriKrushn***a) Agnya deichhanti Brahma Shrimukhare (100)

Table No. 4

Character	**Agent**	**Object**	**Experiencer**	**Addressee**	**Possessor**	**Verb of Addressing**	**Analysis Line**
Sahadeva	Sahadeva		Prabhum (***SriKrushn***a)	Sahadeva (Addressee verb 'Bachane')	Mana	Chinta Laagibaa	Line -1
SriKrushna	***SriKrushn***a	-	Deva Digapaala	***SriKrushn***a		Sumariba	Line -2
Deva Digapaala	***SriKrushn***a	Deva Digapaala	Naarayana (***SriKrushn***a)	***SriKrushn***a	-	Drushya Heba	Line -3
Deva Digapaala	Deva Digapaala	-	Naarayana (***SriKrushn***a)	-	-	Milibaa	Line -4