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

Shagun Sinha
Girish Nath Jha

May 24, 2020
The 6 questions

- What?
- Who?
- How?
- So What?
- Why do we care?
- So What?
Objective

- To study all prospective approaches
- Perform preliminary data study
- Decide which ones to use
The 6 questions

What?

Who?

How?

So What?

Why do we care?

So What?
Text Summarization:
“Distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks)” (Mani and Maybury, 1999)
• Summary:
  “...reductive transformation of source text to summary text through content reduction by selection and/or generalization on what is important in the source”
  (Jones, 1999)
• Types:
  ○ Extractive vs Abstractive
  ○ Single vs Multi document
  ○ Textual vs Multimedia
• Extractive
  ○ Important sentences selected from within the text and quoted verbatim as the summary
  ○ Advantages:
    ■ Summary is good for reference,
    ■ Easy to develop
  ○ Problems:
    ■ Incoherent summaries,
    ■ Unusable summaries
● Abstractive
  ○ Important information selected from text
  ○ Summary produced using new words
  ○ Advantages: Usable, readable summaries
  ○ Problem: Difficult to develop

Point of our focus
The 6 questions

What?
Who?
How?

So What?
Why do we care?
So What?
So What?

- Internet Boom
- Large texts available
- Read more in less time
- Decide whether to read it or not
The 6 questions

What?

Who?

So What?

Why do we care?

So What?

How?
Who?

1958

H. P. Luhn

Long Scientific Paper

Short Abstract
Who?

- Efforts in Indian languages too
- Hindi (Subramaniam and Dalal, 2015)
- Bangla (Sarkar 2012)
- Kannada (Embar et al 2013, Kallimani 2014)
- Mostly extractive
The 6 questions

What?

Who?

How?

So What?

Why do we care?

So What?
Why do we care?

- Scientific literature in Sanskrit
- Manuscript abundance
- Improving semantic processing for Sanskrit
Why do we care?

Efforts in Sanskrit Extractive Text Summarization

- One attempt - Barve et al. (2015)
- Vector Space Model, Inverse Sentence Frequency, Graph-Based Approach
- VSM reported to be successful of all
Why do we care?

Efforts in Sanskrit Abstractive Text Summarization

- None so far
- Ours is the first attempt
The 6 questions

What?
Who?
How?
So What?
Why do we care?
So What?
How?

- **Three ways:**
  - *Structure vs Semantic* Sunitha et al. (2016)
  - *Corpus Based* Hasler et al. (2003)
How?

- **Structure:**
  - Sentences collected in a predefined structure
  - Summary later
  - Eg: Ontology-based etc
How?

- **Semantic:**
  - Document input
  - Semantic Representation
  - Summary based on representation
  - Graph-based method
How?

- Machine Learning:
  - Supervised: Training with text and its labeled Summaries
  - Unsupervised: Graph-based, again, Vector Space, etc
How?

- Corpus-Based
  - Historically used
  - To study features
  - Mostly extractive
Sanskrit text properties: What intrigues us?

1. Compounds and Sandhis
2. Word Significance
3. Verb usage diversity
1. Newspaper articles selected
2. Short sentences
3. Standard Sanskrit usage
4. Variety very less
CONCLUSION

1. Graph-Based
   To develop graphs on short sentences
   Start with semantic graphs
2. Supervised
   Newspaper data
The 6 questions

- What?
- Who?
- How?
- So What?
- Why do we care?
- So What?
Future course

1. Beginning the work
2. See the possibility of corpora-based approach, requiring annotated data
3. Decide future course of action
4. Cross-lingual summarization
References


References


Thanks to: WILDRE Organizers, Microsoft, School of Sanskrit and Indic Studies JNU, Well Wishers
Questions invited...

Contact: shagunsinha5@gmail.com