Text Summarization: Single & Multi-Documment Summarization

Presented By:
Priyanka Ojha
078MSIISE15
Thapathali Campus, IOE

Outline

- Motivation
- Introduction
- Objectives
- Potential Apllication
- Theoretical Background
- Summarization Methods
- Traditional Approaches to Text Summarization
- Modern Approach to Text Summarization
- Multi Document Summarization

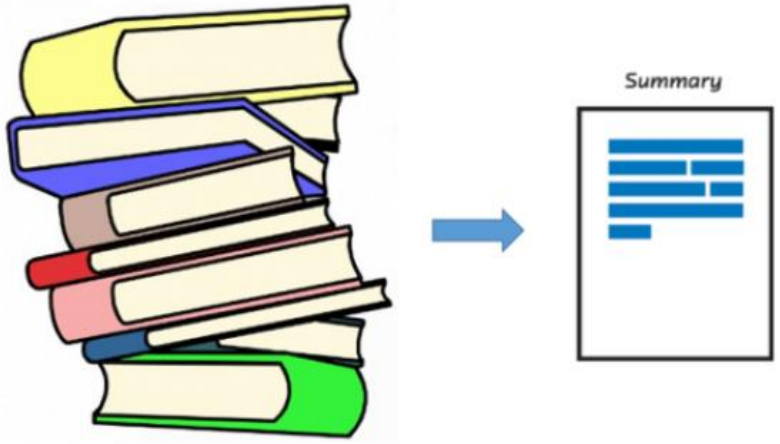
Outline

- Motivation
- Introduction
- Theoretical Background
- Summarization Methods
- Traditional Approaches to Text Summarization
- Modern Approach to Text Summarization
- Multi Document Summarization
- Evaluaion Measures
- Summarization Systems
- Challenges
- Conclusion

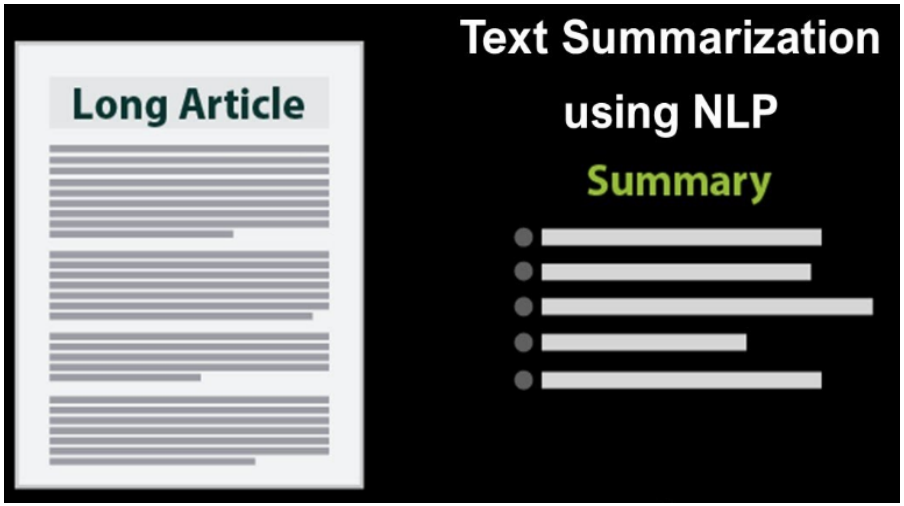
Motivation

- Download 1000+ papers and get the Summary
- You have list of emails about sport event get the summary of those emails in one paragraph
- You have to study loads of books for the exam and the summarizer gives the key concepts of the books as few pages notes

Before



After



Introduction[1]

Automatic Summaries

- Cmputer generated summaries without human intervention
- Should be less than half of the original text
- Should convey important information
- May be produced from single to multiple documents

Objectives

- To reduce a text with a computer program in order to create a summary that retains the most immportant points of the original text.
- To identify the most important information from the given text and present it to the end user

Potential Applications

- Search Engine Optimization
- Summarize of email threads
- Action Item from a meeting
- Simplifying text by compressing sentences

Theoretical Background [1]
Types of Summaries

1. With respect to content
   - Indicative: Provide an idea what the text is about, but do not render the content
   - Informative: Shortened version of the text
2. With respect to way of creating
   - Extract: Identify important section of the text
   - Abstract: Produce important material in a new way
3. With respect to input
   - Restricted vs. Unrestricted domain
   - Single-document vs. Multi-document
4. With respect to Purpose
   - Generic vs. Query based
   - Background vs. Just-the -news

Theoretical Background [2]
Aspects that Describe Summaries[1]

- **Input**
  - Single-document vs. multi-document...fuse together texts?
  - Domain-specific vs. general...use domain-specific techniques?
  - Genre...use genre-specific (newspaper, report…) techniques?
  - Scale and form…input large or small? Structured or free-form?
  - Monolingual vs. multilingual...need to cross language barrier?

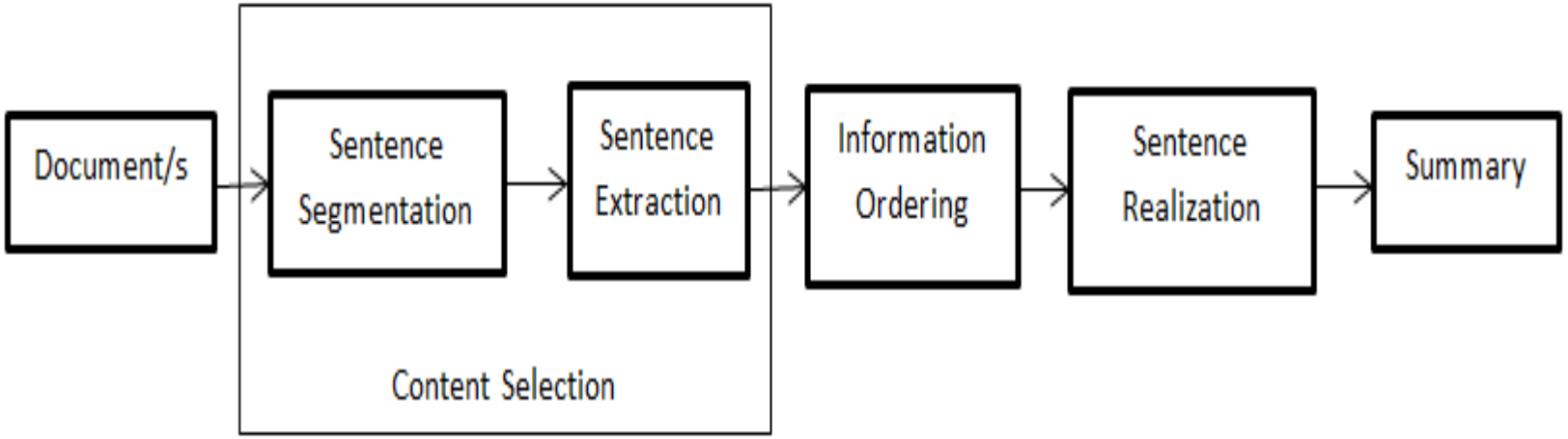
Theoretical Background [3]
Aspects that Describe Summaries[2]

- **Purpose**
  - Situation...embedded in larger system (MT, IR) or not?
  - Generic vs. query-oriented...author's view or user's interest?
  - Indicative vs. informative...categorization or understanding?
  - Background vs. just-the-news…does user have prior knowledge?

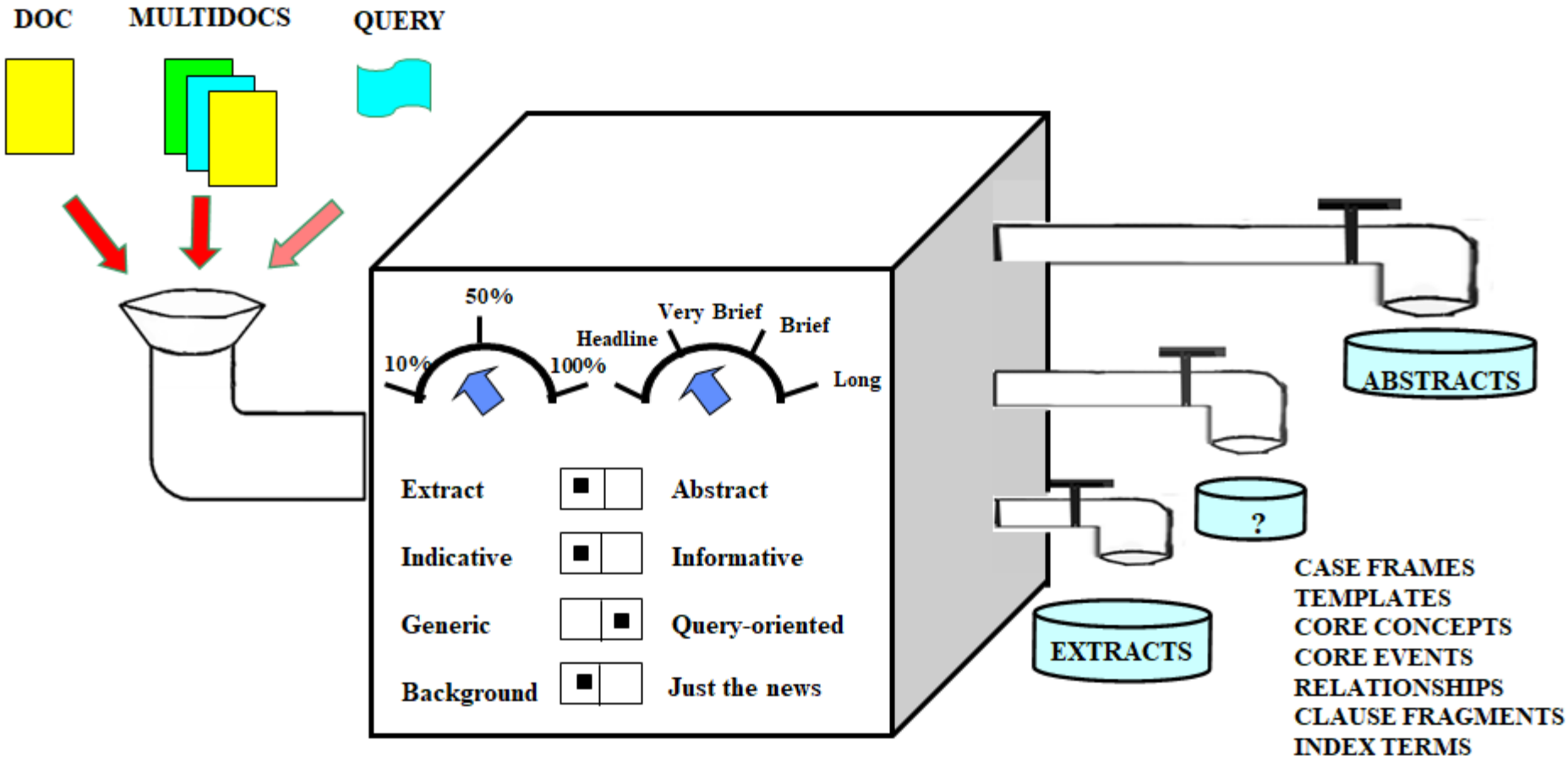
Theoretical Background [4]
Aspects that Describe Summaries[3]

- **Output**
  - Extract vs. abstract...use text fragments or re-phrase content?
  - Domain-specific vs. general...use domain-specific format?
  - Style…make informative, indicative, aggregative, critical...

Work Flow Diagram



Working Principle



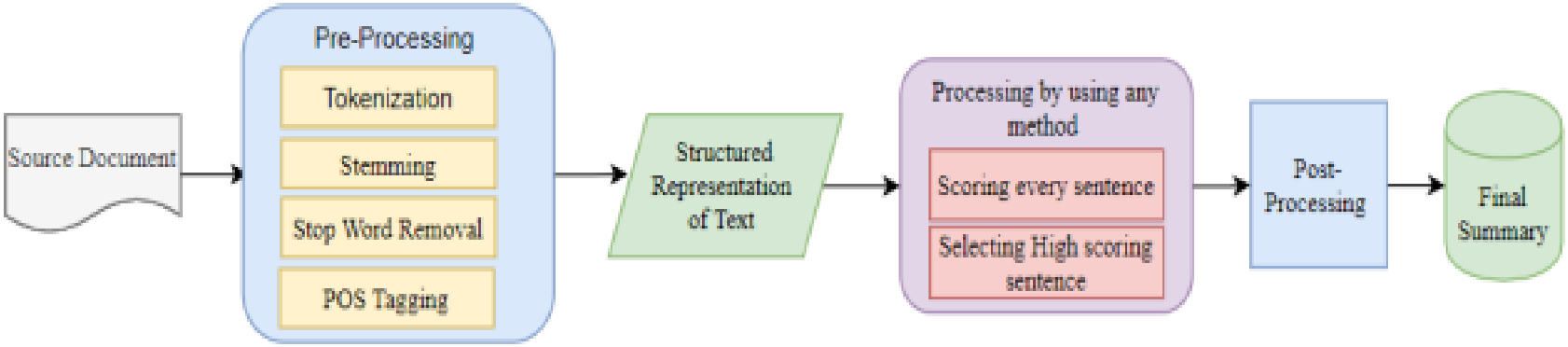
- abc

Summarization Methods

- Extraction:
  - Concatenating important sentences or paragraphs without understanding the meaning of those sentences
- Abstraction:
  - Generating the meaningful summary

Extraction-based Summarization [1]

- Extraction-based summarization techniques produce summaries by choosing a subset of the sentences in the original text



Extraction-based Summarization [2]

- It first creates an intermediate representation for taking out the most important information of the text
- There are two main types of representations:
  - Topic representations: It focuses on representing the topics represented in the texts. Include Latent Semantic Analysis and Bayesian Models.
  - Frequency Driven Approaches: In this approach, we assign weights to the words. If the word is related to the topic we assign 1 or else 0.

Extraction-based Summarization [3]

Topic representation

- It focuses on representing the topics represented in the texts. Include Latent Semantic Analysis and Bayesian Models.
- Latent Semantic Analysis: extracts hidden semantic structures of words and sentences

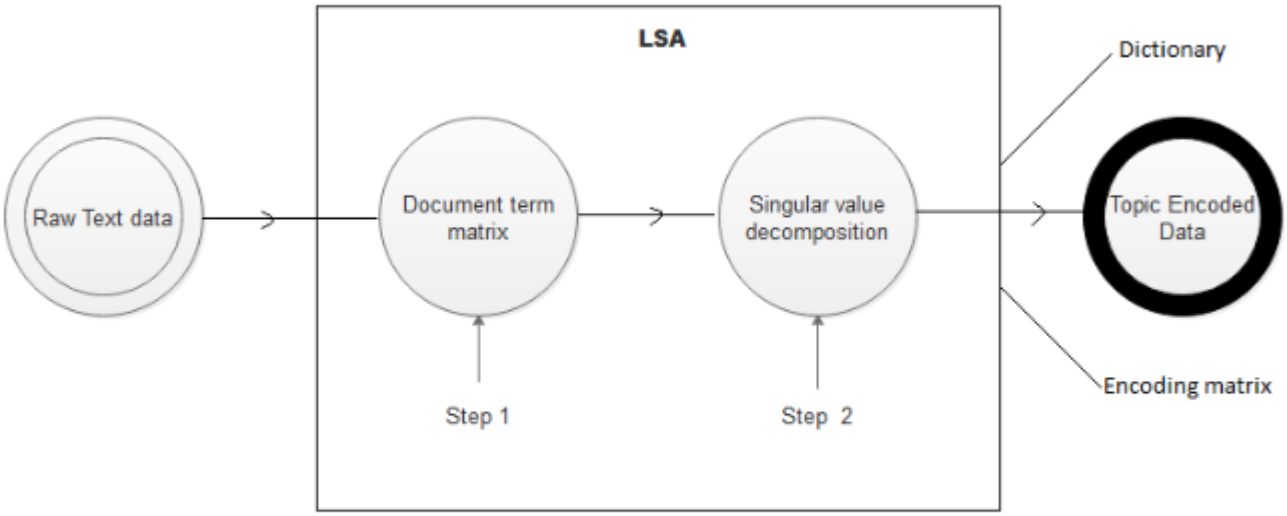


Fig. LSA processing

Extraction-based Summarization [4]

Topic representation

- Bayesian Models:
- It is about ranking sentences according to how useful/important they are as part of summary, we will consider here a particular ranking scheme based on the probability of a sentence being part of summary under a given distribution of votes (DOV)

Extraction-based Summarization[5]
Frequency Driven Approaches[1]

In this approach, we assign weights to the words. If the word is related to the topic we assign 1 or else 0. The weights may be continuous depending on the implementation.

Common techniques for frequency driven approaches are:

- Word Probability:
  - It simply uses the frequency of words as an indicator of the importance of the word.
  - The probability of a word w is given by the frequency of occurrences of the word, f (w), divided by all words in the input which has a total of N words.

Extraction-based Summarization [6]
Frequency Driven Approaches[2]

- TFIDF.(Tern Frequency Invert Document Frequency):
  - Here the TF-IDF method is used for assigning the weights. TFIDF is a method that assigns low weights to the words that occur very frequently in most of the documents under the intuitions that they are stopwords or words like “The”.
  - Otherwise, due to the term frequency if a word appears in a document uniquely with a high frequency it is given high weightage.

Extraction-based Summarization [7]
Frequency Driven Approaches[3]

- Topic word Approaches:
  - It aims to identify words that describe the topic of the input document.
  - This method calculates the word frequencies and uses a frequency threshold to find the word that can potentially describe a topic. It classifies the importance of a sentence as the function of the number of topic words it contains.
  - Indicator Representations, Graph-Based Methods, Machine-Learning Methods

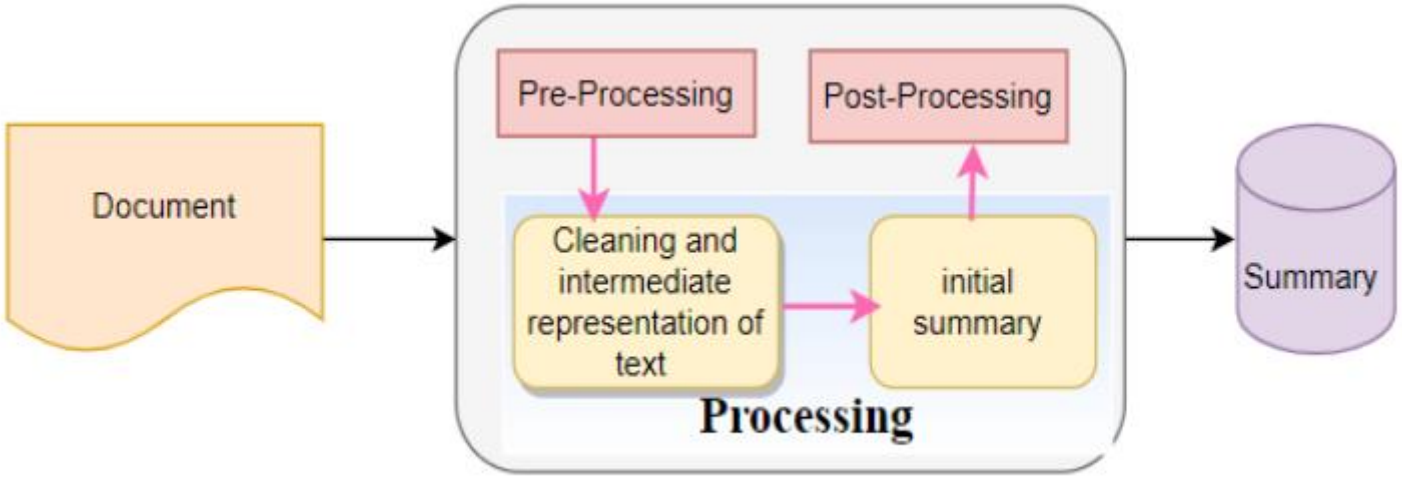
Extraction-based Summarization [8]

- After we get the intermediate representations, we move to assign some scores to each sentence to specify their importance.
- For topic representations, a score to a sentence depends on the topic words it contains, and for an indicator representation, the score depends on the features of the sentences.
- Finally, the sentences having top scores, are picked and used to generate a summary.

Abstraction-based Text Summarization[1]

- It generates new sentences that could best represent the whole text.
- It includes heuristic approaches to train the system in making an attempt to understand the whole context and generate a summary based on that understanding.
- This is a more human-like way of generating summaries and these summaries are more effective as compared to the extractive approaches.

Abstraction-based Text Summarization Methods[2]



Abstraction-based Text Summarization[3] Methods

- Structured Based Approach
  - Tree based method
  - Template based method
  - Ontology based method
  - Lead and body phrase method
  - Rule based method
- Semantic Based Approach
  - Multimodal semantic model
  - Information item based method
  - Semantic Graph Based Method
- Methods based on deep learning

Computational Approach

- Top-down Approach
  - User wants only certain types of info.
  - System needs particular criteria of interest, used to focus search.
  - **Information Extraction task**: Given a form and a text, find all the information relevant to each slot of the form and fill it in.
  - **Summary-Information Extraction task**: Given a query, select the best form, fill it in, and generate the contents.
- Bottom-up approach
  - User wants anything that's important.
  - System needs generic importance metrics, used to rate content.
  - **Information Retrieval task**: Given a query, find the relevant document(s) from a large set of documents.
  - **Summary-Information Retrieval task**: Given a query, find the relevant passage(s) from a set of passages (i.e., from one or more documents).

Comparison Between Information Extraction & Information Retrieval

Information Extraction	**Information Retrieval**
Approach: try to understand text—transform content into 'deeper' notation; then manipulate that.	Approach: operate at word level—use word frequency, collocation counts, etc.
Need: rules for text analysis and manipulation, at all levels.	Need: large amounts of text.
Strengths: higher quality; supports abstracting.	Strengths: robust; good for query-oriented summaries.
Weaknesses: speed; still needs to scale up to robust open-domain summarization.	Weaknesses: lower quality; inability to manipulate information at abstract levels.

Classical Approaches to Text Summarization

- Surface Level
  - Uses shallow features
  - Selectively combines shallow features to extract information
- Entity Level
  - Builds a representation of text
  - Models text entities and their representation
- Discourse Level
  - Models global structure of text and its relations

Surface Level[1]

- Thematic Features:
  - Uses significant word occurence statistics
  - Important sentences that are extracted: Sentences containing words that occur frequently in a text have higher weight than the rest.
- Location:
  - Position of target sentence in the document
  - Lead method: Extracts only the first sentence.
  - Title-based method: Considers that words in the heading or titles are positive relevant to summarization

Surface Level [2]

- Background:
  - Focus on Title, heading, initial part of the text
- Cue words:
  - Focus on phrases that emphassize the sentence such as: "in conclsion", "important", "in this paper", etc.
  - Claim 1: Important sentences contain ‘bonus phrases’, such as significantly, In this paper we show, and In conclusion, while non-important sentences contain ‘stigma phrases’ such as hardly and impossible.
  - Claim 2: These phrases can be detected automatically
  - Method: Add to sentence score if it contains a bonus phrase, penalize if it contains a stigma phrase.

Entity Level [1]

- Similarity:
  - When two words share a common stem and can be extended to phrases and paraphrases
  - Similarity calculated by vocabulary overlap or with linguistic techniques
- Proximity:
  - Refers to distance between text units

Entity Level [2]

- Co-occurence:
  - Meaning units can be related if they occur in common texts
- Thesaural Relationship among words:
  - Synonym, hypernym, meronym(part of relations)

Entity Level [3]

- Coreference:
  - Referreing expressions can be linked so that, coreference chains can be built with coreferring expressions.
- Logical Relations:
  - Agreement, contradiction, entailment, and consistency

Entity Level [4]

- Syntactic Relations:
  - Syntactic rlations based on parse tree
- Meaning Representation-based Relations:
  - Establishing relations between entities in the text
  - E.g. Predicate-argument relations

Discourse Level

- Format:
  - Format of document sch as hpertext markup or document outline
- Threads of Topics:

- Rhetorical Structure of Text:
  - Argumentative or narrative structure
  - To build coherence structure of a text

Current Approach: Hybrid approach

- Combine and integrate techniques
- E.g. Cue Phrase method combined with position and word frequency based method

Multi-Documment Smmarization

- Content Selection:
- Content Filtering:
  - Sentence position, Stigma Words, Maximum Marginal Relevancy,
- Content Presentation:
  - A Buddy System of paired sentences, time annotation and sequence

Multi-Document Summarization

- Multiple Sources of Information
  - Similarity between topics
  - Supplement each other
  - Occasionally contradictory
- Key Tasks
  - Identifying Key concepts across documents
  - Copinng with redundancy
  - Ensuring final Summary is coherent and complete
- Applications
  - News clustering system

Evaluation Measures

- Intrinsic Method
- Extrinsic Method
- Summary Evaluation Environment SEE
- Evaluation Metrics:
  - Recall, Coverage, Retention, and Weighted Retention
  - Precision and Pseudo Precesion

Evaluation Measures

- **ROUGE:**
- It stands for Recall-Oriented Understudy for Gisting Evaluation.
- It is the method that determines the quality of the summary by comparing it to other summaries made by humans as a reference.
- To evaluate the model, there are a number of references created by humans and the generated candidate summary by machine.
- The intuition behind this is if a model creates a good summary, then it must have common overlapping portions with the human references. .

Text Summarization Systems[1]
MEAD

- Features:
  - Centroid based feature
  - Position and overlap with the first sentence
- Procedure:
  - Topic detection and Tracking to identify all the related to an emerging event
  - From each cluster centroid is built
  - For each sentence three values are computed
    - Centroid Score: Measures how close the sentence to the centroid is
    - Position Score: how far is the sentence with respect to the beginning
    - Overlap with the first sentence or title of the document by calculating tf*idf
  - All measures are normalized and too similar sentences are discarded, left are included in the summary

Text Summarization Systems[2]
NeATS[1]

- Specific genre of newpaper news
- Basic architecture: Content Slection, Content filtering and content presentation
- Content Selection:
- identify important concepts mentioned in a document colletion
- Techniques used are: term fequency topic signature or term clustering
- Content Filtering:
  - Three different filters are used: Sentence position, stigma words and redundancy filter

Text Summarization Systems[2]

NeATS[2]

- To ensure coherence of the summary, it outputs the final sentence in their chronological order.
- iNeaATS: an interactive multi-document summarization system

Text Summarization Systems[3]

WebInEssence

- It is a search engine to summarize clusters of related Web pages which provide more contextual and summary information

Architecture:

- Web-Spider:
  - Collects URL from internet
  - Groups the URL into clusters
- Create a multi-document summary from from each cluster using MEAD centroid-algorithm

Text Summarization Systems[4]
GISTExter

- Prodeces single and multi-document extracts and abstracrs by template-driven IE.
- Performs differently depending on working on single document or multi-document summarization.
- For single document:
  - Most relevant sentences are extracted and compressed by rules learnt from a corpus of human written abstracts
- For multi-document:
- Information extraction techniques

Text Summarization Systems[5]
NetSum

- Works on single document
- Produces fully automated single-document on neural nets
- Uses machine learning techniques
- Supervised learning
- Sentences are ranked using RankNet algorithm

References:

- Chin-Yew Lin and Eduard Hovy, "From Single to Multi-document Summarization:A Prototype System and its Evaluation",2002
- Elena Lloret, "TEXT SUMMARIZATION : AN OVERVIEW"
- Ms. Anusha Pai, "Text Summarizer Using Abstractive and Extractive Method", 2014
- Tadashi Nomoto ,"Bayesian Learning in Text Summarization"

Thank You!!!