Information Retrieval & Text Mining

Text Summarization and Simplification

Dr. Iqra Safdar & Farooq Zaman
Information Technology University

Why Text Summarization and Simplification?

- Motivations
  - Long documents are hard to read and time consuming
  - News articles
  - Consume content faster and more efficiently
  - Deaf people
  - Blind people
  - Second language learners
  - People with aphasia

Why Text Summarization and Simplification?

Today we are overwhelmed with huge amount of information

News articles

Emails

Social Media Streams

Scientific Articles

Books

Websites

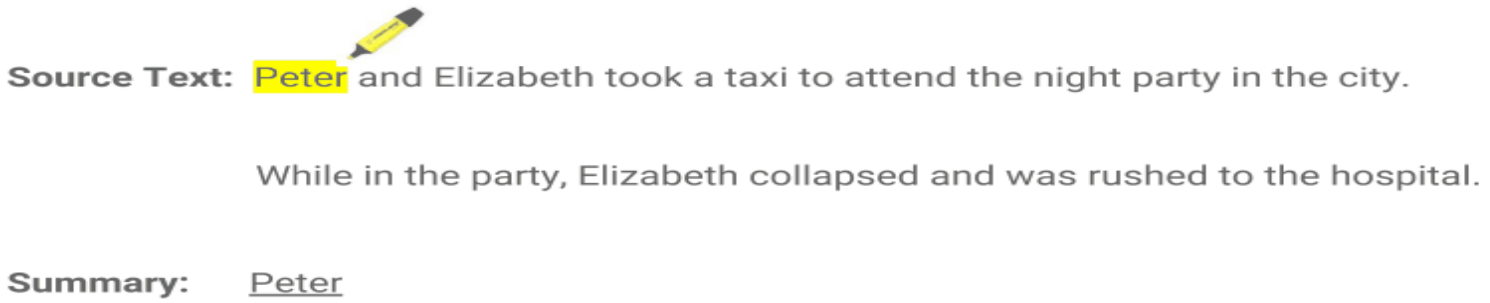
(Nenkova et al., 2011

Text Summarization

- The process of automatically producing short textual document (summary) based on information presented in long textual document (original)

(Collins, Augenstein, & Riedel,
(Nikolov, Pfeiffer, & Hahnlose

- Select important section from original text while ignoring redundant and extra detailed information

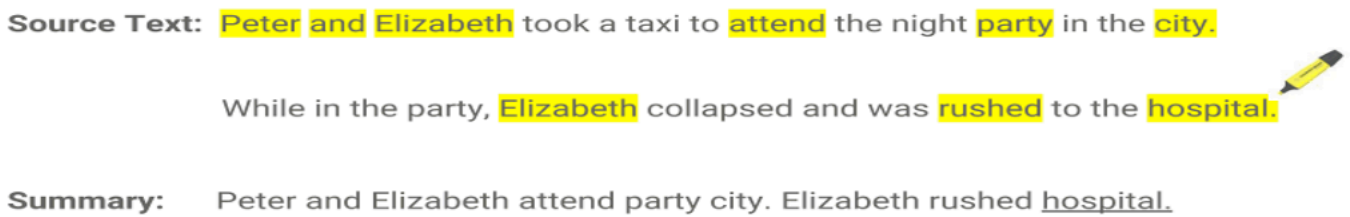


Text Summarization

- The process of automatically producing short textual document (summary) based on information presented in long textual document (original)

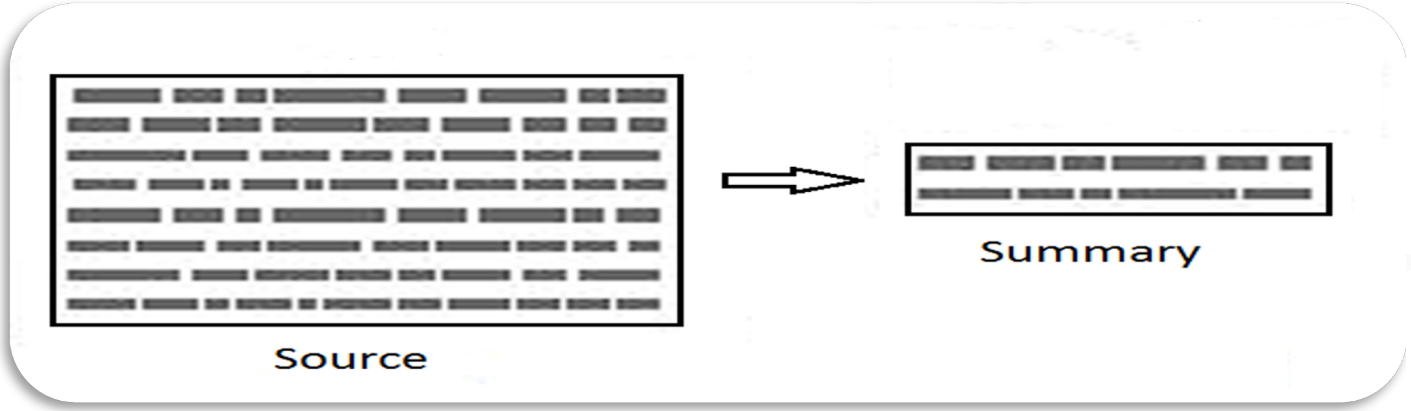
(Collins, Augenstein, & Riedel,
(Nikolov, Pfeiffer, & Hahnlose

- Select important section from original text while ignoring redundant and extra detailed information



Text Summarization (Continue)

- Formally:
  - Given document d consists of words w represents some information I
  - Our goal is to reduce d to d' such that d' consists of w' representing information I'
  - Where $w' \subset w$ and $I' \approx I$

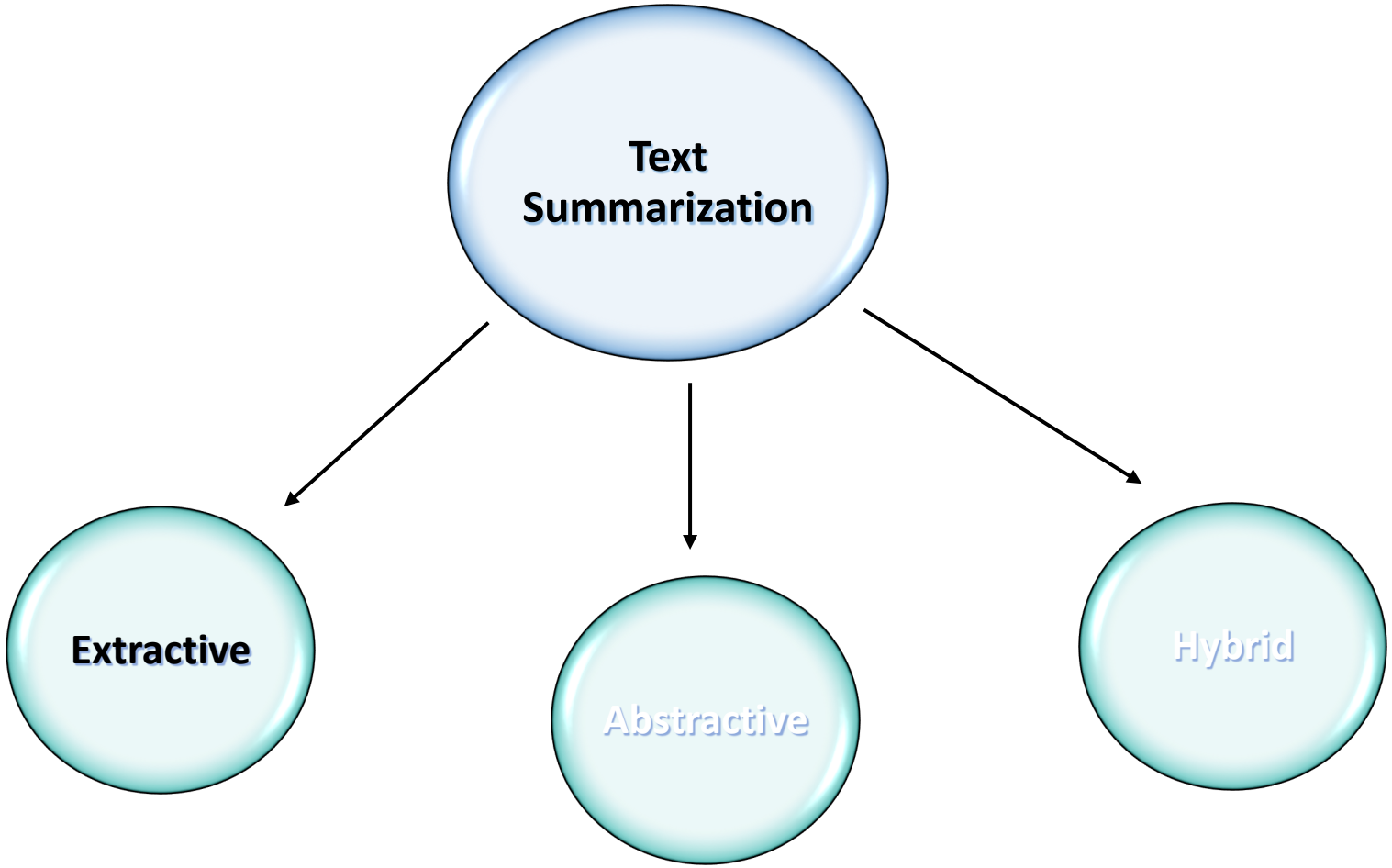


Text Summarization

- The process of automatically producing short textual document (summary) based on information presented in long textual document (original)
- Select important section from original text while ignoring redundant and extra detailed information

- Formally:
  - Given document d consists of words w represents some information I
  - Our goal is to reduce d to d' such that d' consists of w' representing information I'
  - Where $w' \subset w$ and $I' \approx I$

Text Summarization (approaches)



Text Summarization (types)

- Extractive Text Summarization:
  - Select important sentences from original text T and place it in summarize version T' without any modification
  - Need sentence scoring
  - i-e $T' \subset T$

- Abstractive Text Summarization:
  - Generate new text T' based on information presented in original text T
  - May or may not contain any section of the original text T
  - i-e $T \cap T' \geq \emptyset$

Extractive Text Summarization

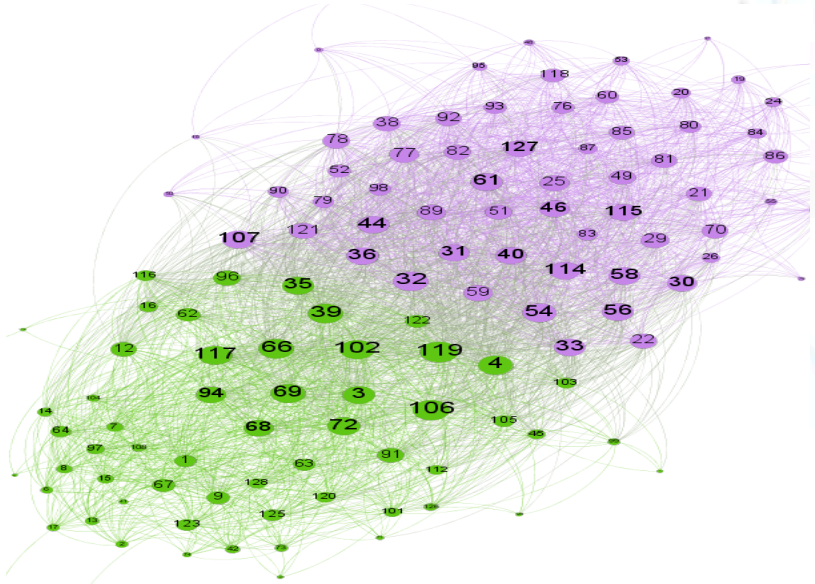
- Steps:
  - Construction of an intermediate representation
    - Topic representation (Signature words)
    - Indicator representation (length of sentence number of key terms)
  - Scoring the sentences (intermediate representation)
  - Selection of a summary
    - Select top K sentences based on computed score

Extractive Text Summarization (cont…)

- Approaches:
  - Topic words (Explanatory words)
    - Number of explanatory words in a sentence
    - Ratio of Explanatory words in a sentence
  - Frequency based approaches (TF-IDF)
  - Latent Semantic Analysis
    - Score Matrix based on TF-IDF
  - Discourse Based Method:
    - Considers connection between sentences
    - Use discourse as atomic unit instead of sentence
  - Indicator representation approaches
    - Represent the text based on a set of features
    - Use these features to directly rank the sentences

Extractive Text Summarization (cont…)

- Approaches:
  - Graph based Methods:
    - Sentences form the vertices
    - Edges between the sentences (similarity between 2 sentences)
    - Cosine similarity with TFIDF
    - Connection to many other sentences in a sub-graph (center of the graph)
    - Independent of language-specific linguistic processing (+)
    - Do not Used syntactic and semantic information (-)
  - Machine Learning:
    - Summarization as a classification problem
      - Naive Bayes
      - Decision trees
      - Support vector machines
      - Hidden Markov models (+)
      - Conditional Random Field (+)



Abstractive Text Summarization

- Approaches:
  - Structure-based approaches
    - Lead and body phrase method
    - Rule-based methods
    - Graph-based methods
    - Ontology-based methods
  - Semantic-based approaches
    - Multimodal semantic model
    - Information item-based methods
    - Semantic Text Representation Model
    - Semantic Graph Model
  - Neural Network Based approaches
    - Encoder decoder RNNs (Generative models)
    - Inspired from Neural Machine Translation

Hybrid Approach

- Combine both Extractive and Abstractive
  - Pointer-generator model
    - Copy words from source texts via a pointer
    - Generate novel words from a vocabulary via a generator
    - Compute the probability of copied word and generated word
    - Use pointing/copying mechanism to generate the output summary

Automated Text Simplification

- Is the process of modifying natural language to achieve reduced complexity and improve readability and understandability

- Approaches
  - Lexical
  - Syntactic
  - Statistical Machine Translation
  - Hybrid Techniques

(Shardlow. M, 2014)

Lexical Approaches

- The process of:
  - Finding complex terms and replace it with simpler terms
  - Explain complicated words expressions by providing definitions/explanations
- Level of granularity
  - Words level
  - Phrase level
- Steps to lexical simplification:
  - Find the complex terms in the text
  - Generate list of substitutions for each complex term
  - Refine the substitutions in step 2 to retain sense in the given context
  - Rank the remaining substitutions according to simplicity
  - Use the simplest synonym as a replacement for the original term

(Stajner & Saggion, 201

Syntactic Approaches

- Syntactic simplification:
  - Is the process of identifying grammatical complexities in a text and rewriting these into simpler structures

- Types of syntactic complexity:
  - Long sentences may be split into their component clauses
  - Sentences which use the passive voice can be rewritten
  - Anaphora may be resolved

Explanation Generation Approach

- Explanation generation
  - Is the technique of taking a difficult concept in a text T and augmenting it with extra information λ which puts it into context and improves user understanding
  - i-e $T' = T + \lambda$
  - Where T' is the output generated simplified text

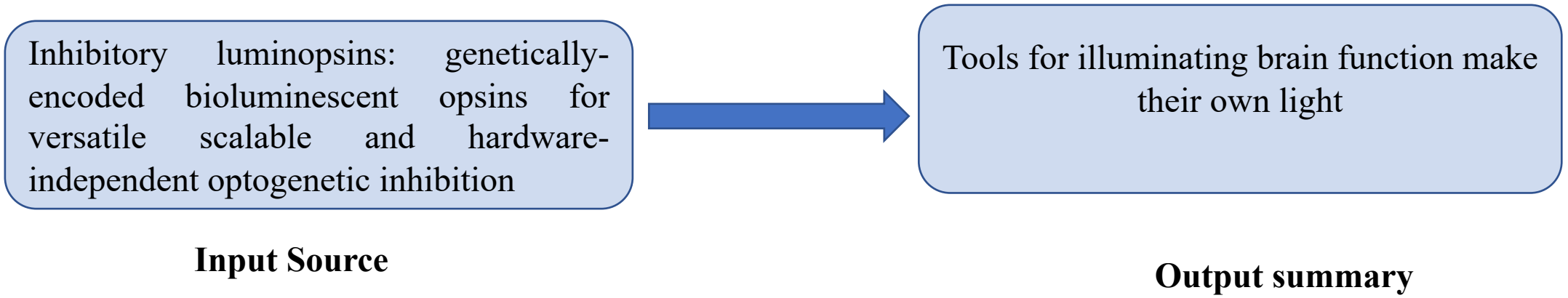
Text Simplification

- The process of substituting difficult complex section of text with easy replacement while maintaining the original meaning of the text
- The simplified document may get increase in length (added explanation)

- Formally:
  - Given document d consists of words w represents some information I
  - enhance d to d' such that d' consists of w' representing information I'
  - Where $w' \supseteq w$ and $I' \approx I$

Text Simplification

- The simplified document may get increase in length (Kriz et al, 2019)



Statistical Machine Translation

- Automated Machine Translation
  - Is an established technique in natural language processing
  - Take input text (source language) transform it into intermediate representation
  - Produce output text (target language) from intermediate representation

- Use case (Text Simplification)
  - Input: Complex text
  - Output: Simple text

Text Simplification (use case)

- Neural Text Simplification of Clinical Letter with a Domain Specific Phrase Table
  - Input: Clinical prescription contain complex clinical terms
  - Output: Patient understandable simple prescription
  - Method: Encoder decoder Recurrent Neural Network with Phrase table

(Shardlow & Nawaz, 2019)

Evaluation

- BLEU (bilingual evaluation understudy):
  - Compare the generated output summary with reference summary produce by human
  - Originally developed for Machine translation systems
  - Can only be applied to Extractive summarization (Lexical overlap)
  - Favor Short sentences (-)

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation)
  - Compare an automatically produced summary against a reference (human-produced)

- METEOR (Metric for Evaluation of Translation with Explicit Ordering)
  - Based on the harmonic mean of unigram precision and recall
  - With recall weighted higher than precision
  - Used for measuring fluency of text

Evaluation (cont…)

- Points to be considered for evaluating TS system
  - Readability
    - Is the generated summary readable ?
  - Grammaticality and meaning preservation
    - Paragraph A is grammatical ?
    - Paragraph B is grammatical ?
    - Paragraphs A & B have the same meaning ?

(Stajner & Saggion, 2018)

References

[1] Shardlow, M. (2014). A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, *4*(1), 58-70

[2] Shardlow, M., & Nawaz, R. (2019, July). Neural Text Simplification of Clinical Letters with a Domain Specific Phrase Table. In Proceedings of the 57th Conference of the Association for Computational Linguistics (pp. 380-389)

[3] Štajner, S., & Saggion, H. (2018, August). Data-Driven Text Simplification. In Proceedings of the 27th International Conference on Computational Linguistics: Tutorial Abstracts (pp. 19-23).