Module 6

Text categorization and summarization

Text Summarization

- Summarization is the task of condensing a piece of text to a shorter version that contains the main information from the original.

- The need for quick acquisition and assimilation of useful insights from a large corpus of information on the Internet has driven the development of various automated summarization systems. These systems provide filtered and high-quality concise content to the users allowing them to work at unprecedented scale and speed.

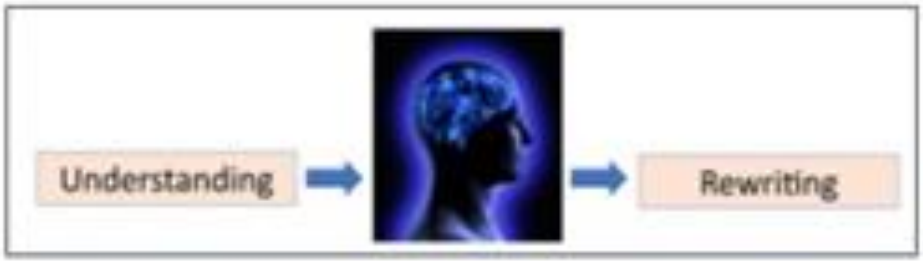
Types of text summarization

- Extraction: - In Extractive text summarization , summary is generated by selecting a set of words, phrases, paragraph or sentences from the original document.
  - Extract certain sentences
  - No issues with grammar
-

What "extraction" of sentences implies?

- Ranking
- Selection

But actual text summarization is much more!



Types of text summarization

- **Abstraction: -** Abstractive methods are based on semantic representation and then use natural language processing techniques to generate a summary that is nearer to summary generated manually.

- This kind of summary may contain words that are not found in the original document. Currently research is going on this method and demand for this method is more.
  - Produce “abstracts”
  - Content Understanding
  - generation

Types of text summarization

- A Survey on Automatic Text Summarization… Dipanjan dasCarnegie Mellon University

requires advanced language generation techniques. In a paradigm more tuned to information retrieval (IR), one can also consider *topic-driven summarization*, that assumes that the summary content depends on the preference of the user and can be assessed via a *query*, making the final summary focused on a particular topic.

Text Summarization

- Summarization categories **based on content of summary:**

- **Indicative Summarization:** It portrays the key topics in the text reducing the length of the original text by 90%, includes metadata like writing style, length of a document, however fails to provide factual information. It helps to decide whether a user wants to read the document or not.

- **Informative Summarization**: It contains content which are generally longer, reduces the length of original text by 70-80%. It includes facts and information which can replace the original text.

Text Summarization

- **Evaluative Summarization:** It aims to capture the opinion or the views of the author on a given topic/subject/product. Sometimes, it is also referred as review or opinion based summarization

Semantic based Abstractive Summarization

- It involves inputting the semantic representation of the document to natural language generation module in order to obtain the desired summary.
- The different methods used in this approach are: (a) Multimodal semantic model, (b) Information item based method, and (c) Semantic Graph based method.
- **Multimodal semantic model** constructs a semantic model by making use of the concepts and finding the relations between these concepts with the helps of an ontology. The next stage involves identifying the important concepts by using information density metrics. The summary is generated from these important concepts in the final stage.

Semantic based Abstractive Summarization

- The **information item based method** first identifies informative items by performing syntactic analysis on the text. Sentences are generated from these items by following the subject-verb-object structure using a sentence generator. The generated sentences are then ranked based on the average Document Frequency score. From this list, the highly ranked sentences are taken to create the summary.
- Finally, the semantic graph approach consists of three phases: (a) representing the entire document by a Rich Semantic Graph (RSG), (b) applying heuristic rules to reduce the complexity of the semantic graph, and (c) generation of abstract summary form the reduced graph.

Challenges

- **Complexity:** Summarization is one of the hardest problem of Natural Language Processing (NLP) as understanding text is very complex. It requires understanding of semantics, inferential interpretations and discourse.

- **Word sense ambiguity:** is ambiguity created as when an abbreviations has more than one acronym. In such case, the acronym should be matched depending on the subject for better understanding.

Challenges

- **Summary evaluation:** it is hard problem. The main hurdle is building a fair gold standard against which the system generated results can be compared. It's also hard to determine what a correct summary is as it is subjective. Further, existing popular evaluation metric (ROUGE score) is not suitable to evaluation of abstract summaries as ROUGE is the measure of n-gram matches. While the abstract summaries can have different words which are not a part of the original source texts.
- **Subjectivity of summary:** summarization quality varies from person to person. one per- son's interest in a body of text is different than another's, resulting in variation in quality of a summary

Challenges

- **High reduction rate:** extracts created from single document summarizer usually aims to have five to thirty percent of the source text's length. However, in the case of multi- document summaries for hand-held devices the compression rates are much lower. Which poses a serious challenge as such high reduction rate is very hard to attain without having expert background knowledge.

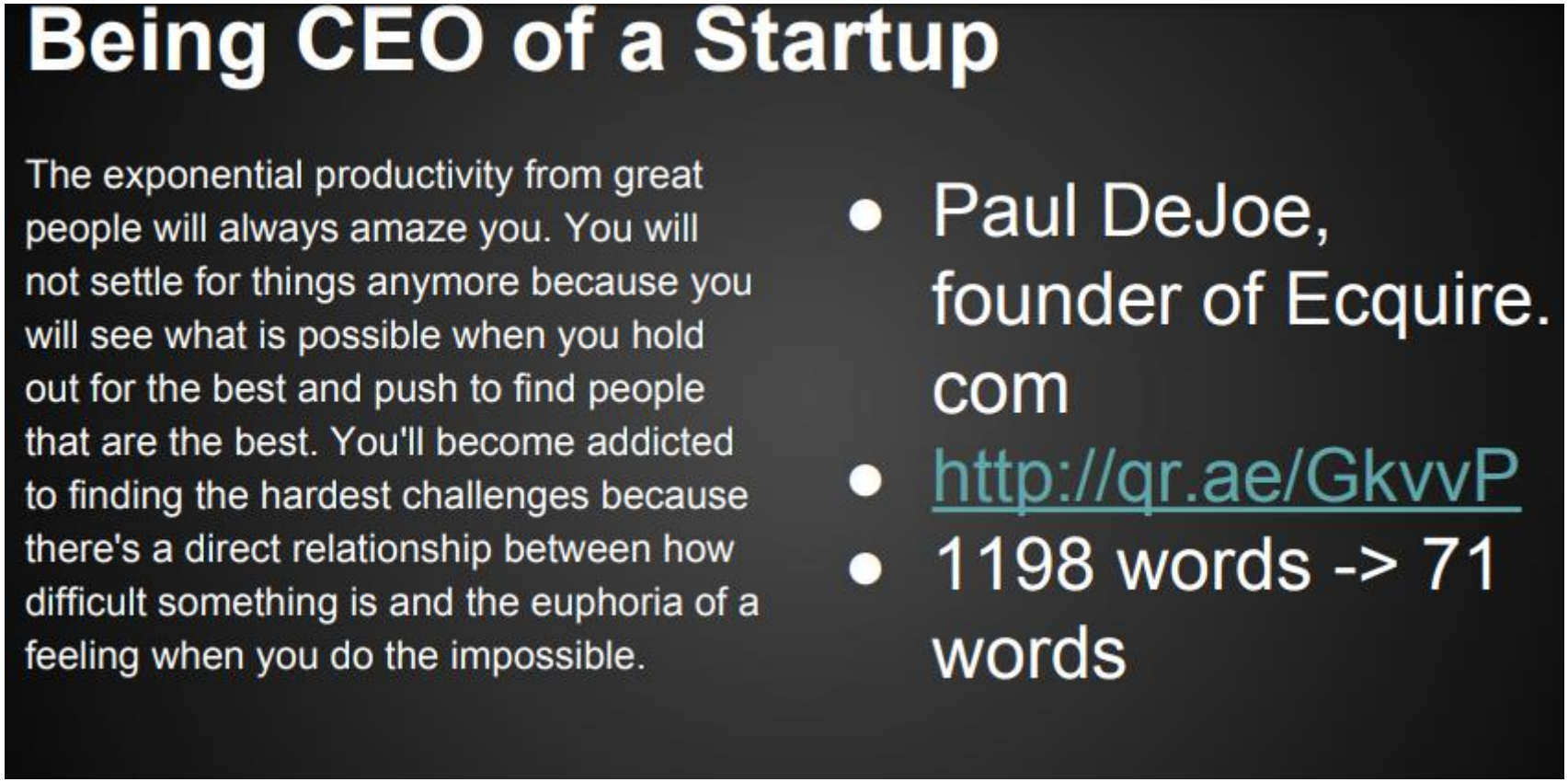
Text Summarization

- While the traditional approach are ruled based and rely mostly on manually compiled features, recent interests has shifted towards using artificial neural network approaches which does not rely on manually compiled features and are rules independent.

Reduction algorithm working

1. Split input text into Paragraph.
2. Split paragraph into sentences.
3. Split sentences into words.
4. Calculate the intersection between 2 sentences.
5. Remove non-alphabetic characters from sentence.
6. Convert content into dictionary.
7. Build the sentence dictionary.
8. Return best sentences in a paragraph.
9. Get the best sentences according to dictionary.

Application News Summarization



Social media Summarization

- deals with summarization of social media text such as: tweets, blogs, community forums etc. These summarization systems are built keeping in mind the needs of the user and are dependent on the genre of social media text. Tweets summarization system will be different from blog summarization systems as tweets has short text (140 characters) and are often noisy. While the blog has considerably longer length text with a different writing style.

Summarization

- Domain specific summaries: Summarization systems are often used in generating domain specific summaries. These systems are designed in accordance with the needs of the user for a specific domain.
- **For example:** legal document summarization deals with generating summary out of a legal/law documents, medical report summarization has aim of generating a summary form a patient report history such that it includes all important clinical events in order of timeline.

Summarization of user generated content

- Summarization of user generated content: deals with summarizing user generated contents like youtube comments, reviews of products, opinions etc. compact version of reviews and comment are very helpful in identifying overall sentiment of mass towards a particular product or topic. Which are often used by the consumers as well as the platform itself in recommending products.

- (Wiki)User-generated content used in a marketing context has been known to help brands in numerous ways.

- **It encourages more engagement** with its users, and doubles the likeliness that the content will be shared.

Summarization of user generated content

- **It builds trust** with consumers. With a majority of consumers trusting UGC over brand provided information, UGC can allow for better brand-consumer relationships.
- **It provides SEO Value** for brands. This in turn means more traffic is driven to the brands websites and that more content is linked back to the website.
- **It reassures purchase decisions** which will keep customers shopping. With UGC, the conversion rate increases by as much as 4.6%.
- **It increases follower count** on various social media platforms.