Ref :https://wikidocs.net/72820

Text Summarization

Extractive Summarization

similar to Google's PageRank. It's an unsupervised, extractive summarization technique that ranks these components based on their relevance and connections, allowing for concise, automated summarization of large volumes of text.

Extractive summarization is a method of creating a summary by selecting a few important core sentences or phrases directly from the original text and combining them. Therefore, all sentences or phrases in a summary produced by extractive summarization are entirely present in the original text. A representative algorithm for extractive summarization is the machine learning algorithm **TextRank**.

3 steps :

1 **Importance Measurement:** This is the phase where the importance of each sentence or phrase is evaluated and scored.

2. **Sentence Selection:** In this step, sentences to be included in the summary are chosen based on their measured importance.

3. **Summary Generation :**This final step involves arranging the selected sentences in their original order to create the final summary.

Abstractive Summarization

Abstractive summarization is a method of summarizing a text by generating new sentences that reflect the core context, even if these sentences were not present in the original text. This approach mimics how humans summarize, and naturally, it is more challenging than extractive summarization. This method primarily uses artificial neural networks, with seq2seq being a representative model.
A drawback of this method is that artificial neural networks like seq2seq are fundamentally supervised learning models. This means that to train an artificial neural network for abstractive summarization, you need not only the 'original text' but also 'actual summaries' as labeled data. Therefore, the process of constructing the data itself can be a burden

3 steps :

1 **Encoding :** This is the phase where the importance of each sentence or phrase is evaluated and scored.

2. **Decoding** In this step, sentences to be included in the summary are chosen based on their measured importance.

3. **Refinement :**This final step involves arranging the selected sentences in their original order to create the final summary.

Advantages

- **Preservation of Original Accuracy:** Since original sentences are used as-is, there is a lower chance of content distortion or misinterpretation.
- **Faster Processing Speed:** It is computationally less expensive and faster than methods that generate new sentences (abstractive summarization).
- **Objectivity:** It is relatively more objective than human intervention in rephrasing content.
- **Connection to Original Content:** It is easy to ascertain which part of the original document the summarized content originated from.

Disadvantages

- **Disjointed Context:** Extracted sentences may not always flow naturally, potentially making the summary awkward or breaking the context.
- **Potential Inclusion of Unnecessary Information:** A sentence deemed important might still contain detailed information irrelevant to the summary.
- **Inability to Create New Information or Creativity:** It is difficult to provide new insights or concise summaries that are not present in the original text.
- **Possibility of Missing Key Information:** Important content might be distributed across multiple sentences or fail to meet the extraction criteria, leading to its omission.

Advantages

- **Higher Fluency and Coherence:** Summaries are often more natural-sounding, grammatically correct, and easier to read because they are generated from scratch.
- **Conciseness:** It can condense information more effectively by rephrasing long sentences into shorter, more direct ones.
- **Novelty and Insight:** Can synthesize information from different parts of the document and present it in a new, insightful way that wasn't explicitly stated in any single original sentence.
- **Handles Complex Relationships:** Better at capturing complex relationships and nuances in the text.
- **Avoids Redundancy:** Can often eliminate redundant information more effectively than extractive methods.

Disadvantages

- **Higher Complexity and Computational Cost:** Requires more sophisticated models and significantly more computational power for training and inference.
- **Prone to Factual Inaccuracies (Hallucination):** Because the model generates new text, there's a risk of "hallucinating" information that is not present in the original document, or misinterpreting facts. This is a major challenge.
- **Difficult to Control Output:** It can be harder to control exactly what information is included or excluded, or to ensure specific keywords are present.
- **Requires Large Datasets:** Training effective abstractive summarization models requires massive datasets of document-summary pairs.
- **Less Transparent:** It's often harder to trace back specific parts of the summary to their original source in the document, making it less transparent.

Key Algorithms and Techniques

- **TextRank / LexRank:** These generate a graph where sentences are nodes and the similarity between sentences are edges, then calculate the importance of sentences in a manner similar to PageRank.
- **KL-Sum:** This selects sentences such that the topic distribution covered by the summary is similar to the topic distribution of the original document.
- **Lead-based Summarization:** This extracts the first few sentences of a document, assuming they are the most important. (The simplest form of extractive summarization.)
- **Luhn's Algorithm:** This identifies key sentences based on word frequency and distribution.
- **Deep Learning-based Models:**
- **Transformer-based models like BERT, RoBERTa:** These understand the meaning of sentences through sentence embeddings and are trained to predict the probability of each sentence being included in the summary.
- **Pointer-Generator Networks (modified for extractive summarization):** These combine extraction and generation, allowing them to either point to existing words or generate new ones. (Strictly speaking, not purely extractive summarization, but includes an extraction mechanism.)
-

Application Areas

- **News Summarization:** Useful for quickly grasping the core content of long news articles.
- **Document Summarization:** Used to identify the key information in extensive documents such as research papers and reports.
- **Information Retrieval Systems:** Generates snippets (summaries) of search results to help users grasp the content before clicking on a link.
- **Customer Service Chatbots:** Quickly summarizes customer inquiries to help agents understand the core issue.

Key Algorithms and Techniques

- **Sequence-to-Sequence (Seq2Seq) Models:** Early abstractive summarization models used Encoder-Decoder architectures, where an encoder processed the input sequence and a decoder generated the output sequence.
- **RNNs (Recurrent Neural Networks) like LSTMs and GRUs:** Historically used as the building blocks for encoders and decoders.
-
- **Attention Mechanisms:** A crucial development that allowed the decoder to "pay attention" to different parts of the input sequence when generating each word of the summary, significantly improving performance.
- **Transformer Architecture:** The current state-of-the-art. Transformers, with their self-attention mechanisms, have revolutionized abstractive summarization. They can process inputs in parallel, making them much faster and more effective than RNNs.
- **BART (Bidirectional and Auto-Regressive Transformers):** A popular pre-trained model for abstractive summarization, often fine-tuned on summarization datasets.
- **T5 (Text-to-Text Transfer Transformer):** Another powerful model that frames all NLP tasks, including summarization, as a text-to-text problem.
- **Pegasus:** Specifically designed for abstractive summarization, often achieving state-of-the-art results.
-
- **Pointer-Generator Networks:** While often used in abstractive settings, they incorporate a "pointing" mechanism (like in extractive) to copy words directly from the source text. This helps prevent out-of-vocabulary words and reduces hallucination, bridging the gap between purely extractive and purely abstractive methods.
- **Reinforcement Learning (RL):** Sometimes used to fine-tune abstractive models, especially for metrics like ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores, by rewarding summaries that are more accurate or relevant

Use Cases

- **News Article Summarization:** Generating concise summaries for news feeds or mobile apps.
- **Meeting Minutes Generation:** Summarizing long meeting transcripts into key action items and decisions.
- **Medical Report Summarization:** Condensing patient notes or research papers.
- **Legal Document Summarization:** Creating overviews of complex legal texts.
- **Customer Service Call Summarization:** Quickly summarizing long customer interactions for agents.