ASSIGNMENT I

Well-posed learning for Automatic Text Summarization Problem

Step 1: What is the Problem?

Task (T): reducing the size of the initial text while at the same time preserving key informational elements and the meaning of content.

Experience (E): knowledge of particular language & able to find importance of sentences or word.

Performance (P): whether the generated summary text covered all important key informational elements and the meaning of generated content will be same as given paragraph.

Step 2: Why does the problem need to be solved?

Motivation

Since manual text summarization is a time expensive and generally laborious task, the automatization of the task is gaining increasing popularity and therefore constitutes a strong motivation for academic research.

Solution benefits

There are important applications for text summarization in various NLP related tasks such as text classification, question answering, legal texts summarization, news summarization, and headline generation. Moreover, the generation of summaries can be integrated into these systems as an intermediate stage which helps to reduce the length of the document.

Solution use

In the big data era, there has been an explosion in the amount of text data from a variety of sources. This volume of text is an inestimable source of information and knowledge which needs to be effectively summarized to be useful. This increasing availability of documents has demanded exhaustive research in the NLP area for automatic text summarization. Automatic text summarization is the task of producing a concise and fluent summary without any human help while preserving the meaning of the original text document.

Step 3: How would I solve the problem?

It is very challenging, because when we as humans summarize a piece of text, we usually read it entirely to develop our understanding, and then write a summary highlighting its main points. Since computers lack human knowledge and language capability, it makes automatic text summarization a very difficult and non-trivial task.

Various models based on machine learning have been proposed for this task. Most of these approaches model this problem as a classification problem which outputs whether to include a sentence in the summary or not. Other approaches have used topic information, Latent Semantic Analysis (LSA), Sequence to Sequence models, Reinforcement Learning and Adversarial processes.

In general, there are two different approaches for automatic summarization: **extraction** and **abstraction**.

The extractive approach

Extractive summarization picks up sentences directly from the document based on a scoring function to form a coherent summary. This method work by identifying important sections of the text cropping out and stitch together portions of the content to produce a condensed version.

A typical flow of extractive summarization systems consists of:

1. Constructs an intermediate representation of the input text intending to find salient content. Typically, it works by computing TF metrics for each sentence in the given matrix.

2. Scores the sentences based on the representation, assigning a value to each sentence denoting the probability with which it will get picked up in the summary.

3. Produces a summary based on the top k most important sentences. Some studies have used Latent semantic analysis (LSA) to identify semantically important sentences.

Abstractive summarization methods aim at producing summary by interpreting the text using advanced natural language techniques in order to generate a new shorter text — parts of which may not appear as part of the original document, that conveys the most critical information from the original text, requiring rephrasing sentences and incorporating information from full text to generate summaries such as a human-written abstract usually does. In fact, an acceptable abstractive summary covers core information in the input and is linguistically fluent.

Thus, they are not restricted to simply selecting and rearranging passages from the original text.

Abstractive methods take advantage of recent developments in deep learning. Since it can be regarded as a sequence mapping task where the source text should be mapped to the target summary, abstractive methods take advantage of the recent success of the sequence to sequence models. These models consist of an encoder and a decoder, where a neural network reads the text, encodes it, and then generates target text.

Which approach to be select?

The extractive approach is easier because copying large chunks of text from the source document ensures good levels of grammaticality and accuracy. On the other hand, sophisticated abilities that are crucial to high-quality summarization, such as paraphrasing, generalization, or the incorporation of real-world knowledge, are possible only in an abstractive framework. Even though abstractive summarization is a more challenging task, there has been a number of advances so far, thanks to recent developments in the deep learning area.