Capstone Project Proposal about Text Summarization using Machine Learning

Domain background

With the rise of information technologies, globalization and Internet, an enormous amount of information is created daily, including a large volume of written texts. The International Data Corporation (IDC) projects that the total amount of digital data circulating annually around the world would sprout from 4.4 zettabytes in 2013 to hit 180 zettabytes in 2025. Dealing with such a huge amount of data is a challenging problem where automatization techiniques can help many industries and businesses. Without summaries it would be practically impossible for human beings to get access to the ever growing mass of information available online. For example, hundreds or thousands of news are published around the world in a few hours and condensing them to make them available to the general public is a manual and very expensive task, in time and money. And people expect to get the information as soon as possible and they only pay attention for a fews seconds, they do not want to read a full article for ten minutos. So the development of automatic techniques to get short, concise and understandable summaries would be of great help to many global companies and organizations. Furthermore, applying text summarization reduces reading time, accelerates the process of researching for information, and increases the amount of information that can fit in an area.

But this is not a recent subject, for decades studies and projects have been carried out in this field in order to achieve automatic text summarization systems. For example, Luhn in 1958 [2] or the DimSum in 1997 [3] are two examples of papers where Natural Language Processing techniques are applied to text summarization. In the 80s and 90s the classical NLP techniques where the foundations of the automatic text summarization systems but recently the advances in deep learning algorithms has become a new source of inspiration to face and solve this problem.

Problem description and statement

Text Summarization is a challenging problem these days and it can be defined as a technique of shortening a long piece of text to create a coherent and fluent summary having only the main points in the document. But, what is a summary? It is a *text that is produced from one or more texts, that contains a significant portion of the information in the original text(s), and that is no longer than half of the original text(s)* [1]. So given a long, multisentence document we need to extract the main concepts, ideas or topics in the document and generate a new short text containing the same concepts, ideas or topics. *Summarization clearly involves both these still poorly understood processes, and adds a third (condensation, abstraction, generalization)*. It is not a well-defined with just one solution problem, there could be many diferent summaries for a given text, so it is not an easy and simple problem. In fact, today is a very active field of investigation and the state-of-the-art solutions are still not so impressive and accurate than we could expect.

The problem, we are facing in this project, is to produce a summary (about 50-word length) from a single document text much more longer, 300-400 words. It will be a single document summarization (there are some techniques to approach a multidocument summarization but is not in the scope of this paper) and its output could be an *extract* (containing pieces of the source text) or *abstract* (a new text is created). And we are looking for an efficient method, not very time comsuming, that can be applied easily in a software solution.

Dataset and inputs

When searching for information and data about text summarization I found hard to obtain a "good" dataset. Some of the most popular data sets are intended for research use, containing hundred of thounsands examples and gigabytes of data that require high computational capacity and days or weeks to train. But we were interested in a dataset that could be trained faster, in a few hours, where we can experiment and develop a !!!!!!.

We will use a dataset from Kaggle, here is the link, that consists in 4515 examples and *and contains Author_name, Headlines, Url of Article, Short text, Complete Article*. This data was extracted from Inshorts and scraped the news article from Hindu, Indian times and Guardian. There are two sets of data: one containing some extra features like Author, Date and the other one with only two columns, text and headline (or summary). This last one

An example:

Text: *"Isha Ghosh, an 81-year-old member of Bharat Scouts and Guides (BSG), has been imparting physical and mental training to schoolchildren in Jharkhand for several decades. Chaibasa-based Ghosh reportedly..."*

Summary: *"81-yr-old woman conducts physical training in J'khand schools"*

Use of others dataset????

Solution Statement

Our goal will be to build a machine learning model to produce a short summary from a source text. We will analyze and apply one or two commonly used methods, compare with each other and measure its performance. There are two main approaches to summarization:

- Extraction-based summarization: identify and extract keyphrases from the source, evaluate its *"importance"* and make a summary with the most valuable phrases. The extraction is made according to the defined metric without making any changes to the texts.

Here is an example:
Source text: Joseph and Mary rode on a donkey to attend the annual event in Jerusalem. In the city, Mary gave birth to a child named Jesus.
Extractive summary: Joseph and Mary attend event Jerusalem. Mary birth Jesus.

- Abstraction-based summarization: entails paraphrasing and shortening parts of the source document, new sentences are created. The abstractive text summarization algorithms create new phrases and sentences that relay the most useful information from the original text — just like humans do.

Abstractive summary: Joseph and Mary came to Jerusalem where Jesus was born.

The latest advances in deep learning has been a boost for the development of abstraction-based model but they requiere a lot of data and expensive computational resources so we will try to compare both methods, evaluate them and choose the one with an appropiate trade-off between performance and resources-consume.

To evaluate the model we will use the ROUGE metric, *"ROUGE, or Recall-Oriented Understudy for Gisting Evaluation,[4] is a set of metrics and a software package used for evaluating automatic summarization and machine translation software in natural language processing. The metrics compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation."*. There are some variations of this metric: ROUGE-N, ROUGE-S and ROUGE-L can be thought of as the granularity of texts being compared between the system summaries and reference summaries. For example, ROUGE-1 refers to overlap of unigrams between the system summary and reference summary. ROUGE-2 refers to the overlap of bigrams between the system and reference summaries.

Bechmark Model

The porpuose of this project is not to define a model to become a state-of-the-art solution to this problem because it will requiere a long investigation and a very large dataset to train for many days. We want to solve the problem of extracting a short summary from a dataset of news

Content

- Predicting_Mortgage_Approvals_EDA
  Code and visualization in Python to develop a Exporatory Data Analysis of the problem and data.

Machine learning model built in Azure Machine Learning Studio

From Microsoft Doc:

*Microsoft Azure Machine Learning Studio (classic) is a collaborative, drag-and-drop tool you can use to build, test, and deploy predictive analytics solutions on your data. Azure Machine Learning Studio (classic) publishes models as web services that can easily be consumed by custom apps or BI tools such as Excel.

Machine Learning Studio (classic) is where data science, predictive analytics, cloud resources,

and your data meet.*

In the second part of the capstone I built a predictive model on Azure ML Studio to predict when an mortgage approval would be accepted or not, here is the link for a description

Links related:

- https://towardsdatascience.com/a-quick-introduction-to-text-summarization-in-machine-learning-3d27ccf18a9f

[1] - Hovy, E. H. Automated Text Summarization. In R. Mitkov (ed), The Oxford Handbook of Computational Linguistics, chapter 32, pages 583–598. Oxford University Press, 2005
[2] - Luhn, H., P. The Automatic Creation of Literature Abstracts. In Inderjeet Mani and Mark Marbury, editors, Advances in Automatic Text Summarization. MIT Press, 1999
[3] - Aonet, C., Okurowskit, M. E., Gorlinskyt, J., et al. A Scalable Summarization System Using Robust NLP. In Proceedings of the ACL'07/EACL'97 Workshop on Intelligent Sclalable Text Summarization, pages 66-73, 1997
[4] - Lin, Chin-Yew. 2004. ROUGE: a Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, July 25 - 26, 2004.

License