Automatic summarization

A Orynbayeva

Faculty of Information Technology, Kazakh-British Technical University, Tole Bi Str. 59, 00500 Almaty, Kazakhstan

**Corresponding author's e-mail: altynai_56@mail.ru*

Abstract

The volume of accessible data on the Web is expanding very quickly. The necessity of frameworks that able to work with those amount of information is becoming ever more desirable. Therefore, dealing with them, it can be beneficial to work with one field of natural language processing called summarization. Automatic summarization plays crucial role in compacting expansive amounts of text into effective summary. This study means to investigate some of the most relevant approaches both in the areas of single-document and multiple-document summarization, giving particular underline to some methods and extractive strategies.

Keywords: natural language processing, automatic summarization, single-document, multiple-document

1 Introduction

Natural language texts are the most common form of knowledge representation, which are easily perceived and interpreted by a human. However, the volumes of these texts have increased significantly and it is not easy to search, process and analyse their contents manually.

As a result, nowadays, there are many technological application which focuses on the analysing and understanding human languages and can be considered by big field named Natural Language Processing (NLP). For instance, entity linking and information extraction, sentiment analysis and opinion mining and also text summarization.

2 Goals

The goals of this study are to study a big area called "Natural Language Processing" and to provide a comprehensive overview of field within NLP named automatic summarization. Moreover, aims of this study to build web-based platform and use known tools and libraries for the reading of articles and documents, whether to compare similarity between articles and their shared keywords, to identify main problem there and to resume it. The idea is to construct the automatic summary from the information by analysis. The service will be built on Djangos framework and will be written with python and their useful library NLTK (Natural language toolkit).

3 Background

Summarization is the technique to reduce a text of document with lines of code based on machine learning or algorithms in order to create a summary which will retain crucial points of the original information. Although research on summarization started approximately 55 years ago, there is still a long trip to research in this field. Over time, attention has drifted from summarizing scientific articles, news, mail messages, blogs and medical sources.

These systems are designed to take information or url of interested website as input and to produce a concise summary of the most related points as output.

There are two methods to automatic summarization: extractive and abstractive ways. Extractive approaches select a subset of existing words or sentences from the original text to form the informative summary, whereas abstraction is important and active research area due to their complexity to research. Furthermore, abstractive method construct semantic representation and generate summary which is closer to human languages. The difficulty differential between these two approaches is greatly increased when the task is handed over to computers. Even with the current state of the art in artificial intelligence, computers are still not nearly advanced enough to support the ability to "reorganize, modify and merge information expressed in different sentences in the input." [4]

Methods of automatic summarization also divide based on the number of sources of information: single-document or multi-document.

4 Surveying the Field of Automatic Summarization

A summary must prioritize the most important themes, sentences. The well-known fundamental method determined by Luhn to identify significance of a sentence. [6] A word's significance is equal to its probability of occurring, which is defined by:

$$\text{significance}(w) = p(w) = c(w)/N, \quad (1)$$

where p(w) is probability of a word, w, occurring and c(w) is number of times a word, w, occurs in the input document and N is total number of words in the input.

```
import requests
    from bs4 import BeautifulSoup
    from nltk.tokenize import word_tokenize, sent_tokenize
    from nltk.corpus import stopwords
    from string import punctuation
    from collections import defaultdict
    from nltk.probability import FreqDist
    from heapq import nlargest
    from nltk.corpus import stopwords



    def getTextTechCrunch(url):
        page = requests.get(url).text
        soup = BeautifulSoup(page , "html.parser")
        article = soup.find_all("div", { "class" : "article-entry text" });
        text = ' '.join(map(lambda p: p.text, article))
        return text


    def summarize(text , n):
        sents = sent_tokenize(text)

        assert n <= len(sents)
        words = word_tokenize(text.lower())
        _stopwords = set(stopwords.words('english') + list(punctuation))

        words = [word for word in words if word not in _stopwords]
        freq = FreqDist(words)

        ranking = defaultdict(int)

        for i, sent in enumerate(sents):
            for w in word_tokenize(sent.lower()):
                if w in freq:
                    ranking[i] += freq[w]

        sents_idx = nlargest(n , ranking , key = ranking.get)
        return [sents[j] for j in sorted(sents_idx)]


articleURL = 'https://techcrunch.com/2017/01/07/using-data-science-to-beat-cancer/'
text = getTextTechCrunch(articleURL)
print(summarize(text , 2))
```

It was one of the approaches to summary extraction. Summary extraction can be frequency based approach, feature based approach and machine learning based approach. Two techniques that use frequency based approach are word probability and term frequency-inverse document frequency.

One of the other way to determine the importance of a sentence is based on feature approach which reflects the relevance of that sentence that can be shown from sentence position, presence of title word and keywords.

In later works from journal od Computer Science, there was used Particle Swarm Optimisation (PSO) algorithm, genetic algorithm, differential evolution algorithm and fuzzy logic in order to enhance finding important sentence by combining term frequency weight with position and node weight [9].

Machine Learning (ML) approach need to have a set of training document (dataset) and their corresponding summary extracts. There are some well-known methods such as Naive Bayes Classifier and Markov Hidden Model, also Neural Network.

In using Naive Bayes method, there are given a sentences where the probability being chosen to be included in the summary is:

$$P(s \in S \mid F_1, F_2, \ldots, F_n) = \frac{\prod_{i=1}^{n} P(F_i \mid s \in S) * P(s \in S)}{\prod_{i=1}^{n} P(F_i)}$$

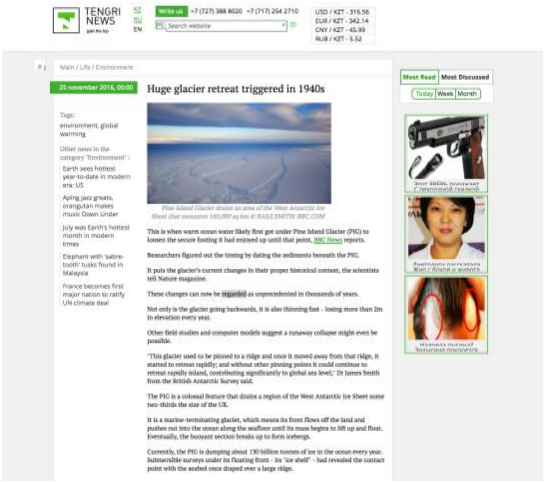
where F1, F2, …, Fn are the sentence features (assuming the features are independent of each other) for the classification and S is the summary to be generated.

Each sentence is then scored according to Equation 2 and ranked for summary selection [7].

Neural network have the advantages to learn summary sentence attributes. The network learn best features and patterns from training to determine the most important information.

5 Simple experiment

In order to better understand how work with NLTK library which is the leading platform for building Python programs to work with human language data [8] and their features, there was done a simple script to summarise the article from Tengrinews (https://en.tengrinews.kz/environment/Huge-glacier-retreat-triggered-in-1940s-263491/) which is kazakh news portal by url getting using BeautifulSoap based on frequency approach to generate summary. As a result, there have been summarised article from Tengrinews into several sentences which more relevant to the summary identifying important sentences. The text was as following: “"This glacier used to be pinned to a ridge and once it moved away from that ridge, it started to retreat rapidly; and without other pinning points it could continue to retreat rapidly inland, contributing significantly to global sea level," Dr James Smith from the British Antarctic Survey said. Currently, the PIG is dumping about 130 billion tonnes of ice in the ocean every year. Submersible surveys under its floating front - its "ice shelf" - had revealed the contact point with the seabed once draped over a large ridge.”



6 Conclusions

In this paper, the research focuses on summary evaluation and the implementation of tools for NLP tasks and especially for automatic summarization. The fundamental concepts and methods related to automatic text summarization have been discussed. Moreover, there was a simple experiment to better understanding working od special python library in order to summarise article from tengrinews with extractive way. It seems that future trend in automatic summarization is not only to focus on the summary information content, however, efforts should also be put into the readability approach of the generated summary and similarity to human language understanding.

References

[1] Jurafsky D, Martin J H 2009 *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics* 2nd edition. Prentice-Hall
[2] Manning C D, Schütze H 1999 *Foundations of Statistical Natural Language Processing* MIT Press
[3] Marcu D 2000 *The Theory and Practice of Discourse Parsing and Summarization* ISBN 0-262-13372-5
[4] Nenkova A, McKeown K 2011 Automatic summarization *Foundations and Trends in Information Retrieval* **5** 103–233
[5] Luhn H P 1958 The automatic creation of literature abstracts *IBM J. Res. Dev.* **2**(2) 159–65 Available: http://dx.doi.org/10.1147/rd.22.0159
[6] Edmundson H P 1969 New methods in automatic extracting *J. ACM* **16**: 264-85 DOI: 10.1145/321510.321519
[7] http://www.nltk.org/
[8] *Journal of Computer Science Science publications* Available: http://thescipub.com/journals/jcs