Dublin City University

School of Computing

CA4009: Search Technologies

Section 4: Summarisation

Gareth Jones

October 2019

Introduction

Simply requiring the searcher to read whole documents when looking for relevant information among retrieved items will often be very inefficient.

The reader may spend much time navigating their way through material in which they have no interest in order to find material that they are looking for.

A key challenge for a search engine is to enable the searcher to efficiently determine which of the retrieved documents are most likely to address their information need.

In order to assist them in determining document relevance IR systems typically return a simple *summary* (often referred to as a *snippet*) of each returned document.

Introduction

Snippets are typically presented in the ranked order of estimated relevance of the documents that they represent in a Search Engine Results Page or SERP.

Snippet creation is a form of content *summarisation*.

A key challenge in snippet creation is to determine which content to include that is most likely to assist the user in determining the potential relevance of each document.

Similar to information retrieval, summarisation is one of the longest standing areas of investigation in computing dating back to the 1950s.

In this section we will explore the general concepts of summarisation, and then relate these to the use of summarisation for snippet creation in IR.

General Definitions of a Summary

Definition: *Summary*: a condensed derivative of a source text.

i.e. content reduction through *selection* or *generalisation* on what is important in the source.

selection - forming a summary focused on a subset of the topical content of the source document in detail

generalisation - forming a summary which overviews the entire topical contents of the source document

A summary thus has a complex relationship with its source text.

Unsurprisingly, a summary usually has a higher density of words relating to the subject of the document than the original document.

Factors Affecting Summarisation

It is hard to assess whether a summary will meet the needs of the user, partly because it is hard to know what the needs of the user actually are.

Since much information is removed from the original document in creating a summary either by a generalisation or selection process, ...

we need to try to ensure that the right information is retained in the summary to satisfy the needs of the reader.

Factors Affecting Summarisation

In considering the needs of the reader, we need to be aware of the intended purpose of the summary.

- What is the function of the summary, e.g. informing or alerting?
- Is the summary for a narrow targeted or more general audience?
- What level of subject knowledge should be assumed of the reader?
- What is the desired format of the output summary?

Should the summary attempt to cover all material in the document or only specified areas? i.e. is it *generalisation* or *selection*?

Factors Affecting Summarisation

In addition to the issue of determining what the user needs from the summary, we also need to consider the form of the input source text.

- For example, a management plan or a scientific document, or other document type, will have fundamentally different forms.

Summarising a document of a different form poses different challenges and may require different procedures.

Factors Affecting Summarisation

The output of a summariser is influenced by multiple factors:

- The form of the input content, e.g. natural language text, bullet points, tables, etc.
- The purpose of the summary.
- The required form and limitations of the output, e.g. length that can be accommodated, elements (text, tables, graphs, images, etc.), screen space available.

Factors Affecting Summarisation

Since the user needs and form of summaries are so varied, it seems unlikely that we can to expect to develop a single general purpose technology for automatic summarisation.

Instead a range of summarisation techniques have been developed which are suitable for different situations, although even following this strategy effective summarisation still presents challenges.

Human Summarising

Summaries have traditionally been created from source texts by human summarisers, so it is instructive in our pursuit of automatic summarisation to consider the human summarisation process.

Professional human summarisers typically generate their summaries to fit a pre-defined set of guidelines (formal or informal).

These guidelines may describe:

- the desired style of the summary language,
- the degree of reduction in the amount of text,
- intended audience,
- format of the output.

Human Summarising

In human summarising there is a strong emphasis on the purpose of the summary, e.g. abstracting for a scientific journal, creating a news summary for a “broadsheet” newspaper.

There will often be several cycles of reviewing and redrafting of the summary.

The completion of the summary may be guided by a checklist to ensure that the guidelines have been followed and all the requirements for the summary have been considered.

The instructions to professional human summarisers will generally be too abstract for automation.

Automatic Summarisation

Automatic summarisation can be broadly divided into two classes:

Information Extraction and Synthesis

The summary contains information extracted from the original text, but is itself a new document.

Information is extracted from the document using natural language processing methods.

- Information extraction: names of individuals, places, organisations, etc; relationships between entities; actions (*who did what?*, *what happened to whom?*, etc.); ... are placed into a database.

Automatic Summarisation

A new text is synthesised using automatic text generation methods.

- Text synthesis: text structure planning; selection of information from database; natural language text generation.

We will not consider this approach further in this module.

Sentence and/or Phrase Extraction

Summary is composed of a subset of the sentences and/or phrases taken directly from the original document.

Automatic Summarisation

- Generally much shallower (and easier!) in terms of trying to understand the text than the information extraction and synthesis method.
- Strategy:
  - Score all the sentences or phrases (somehow).
  - Use the highest scoring sentences as the summary.

    Perhaps consider context in deciding which ones to use, e.g. is the sentence likely to make sense without the one before it if it starts with a pronoun, e.g. “the”, “she”, “it”.

Summarisation by Sentence/Phrase Extraction

Important sentences and phrases in a document are identified within the document and taken as the summary of the document.

For example:

- Each sentence may be scored using some metric (or more likely a combination of metrics) indicating its importance in the document.
- Top n scoring sentences then selected, or sufficient sentences taken to reduce the document to $m\%$ of its original length.

These summaries can be difficult to read (why?), but should enable the topic of the document to be understood.

Often not sufficiently fluent or comprehensive to replace document, but can indicate whether the original document will be of interest to the user.

Summarisation by Sentence/Phrase Extraction

Possible sentence selection criteria:

Summarisation by Sentence/Phrase Extraction

Example of a system for generating a simple document summary.

Form document summary by selecting most “important” sentences.

Sentences scored using the following factors:

- Luhn’s score for clusters of significant words.
- Frequency of document title words in the sentence.
- Location of sentence within the document.
- Frequency of query words in the sentence.

The following simple summarizer was designed experimentally, and was originally described by Tombros and Sanderson (ACM SIGIR 1998) .

Luhn's Keyword Cluster Method

- As noted in the earlier discussion on term proximity for IR, in addition to his work on term frequency in a document, Luhn also observed that the relative location of words in a document indicates their probable relationship to each other.
- *Significant* words occur between low and high frequency limits.
  - Stop words are ignored.
  - Very rare words can also be ignored - $n(i)$ has a very small value.
- Luhn determined that two significant words are significantly related if they are separated by not more than five insignificant words.
- This rule can be used to identify significant clusters of words in a sentence.

Luhn's Keyword Cluster Method

The procedure for locating word clusters operates as follows:

- Find the first significant word in the sentence.
- Locate the last significant word before the sentence ends or there is a sequence of five non significant words.
- Bracket the phrase with the first and last significant words at the end.
  e.g. "The sentence [**scoring** process utilises **information** both from the **structural**] organization."
- Calculate the significance score of the sentence using the following equation.

Luhn's Keyword Cluster Method

Luhn's cluster significance score factor for a sentence is given by,

$$SS1 = \frac{SW^2}{TW}$$

where $SS1$ = sentence score

SW = number of bracketed *significant* words (in this case 3)

TW = total number of bracketed words (in this case 8)

Thus $SS1$ for the above sentence is 1.125.

If two or more clusters of significant words appear in a given sentence - the one with the highest score is chosen as the sentence score.

Title Terms Frequency Method

- The title of an article often reveals the major subject of that document.
- Each sentence containing one or more of the title terms can be considered to be more significant in the document.
- For each sentence a title score can be computed as follows,

$$SS2 = \frac{TTS}{TTT}$$

where $SS2$ = title score for a sentence

TTS = total number of title terms found in a sentence

TTT = total number of terms in a title

Title Terms Frequency Method

TTT is a normalization factor for the score contribution of $SS2$.

Without TTT, $SS2$ could come to dominate the overall sentence score - see later.

Location/Header Method

- The first sentences of a document and section headings often provide important information about the content of the document.
- The first few sentences and the section headings of an article can be assigned a location score to boost their overall score as follows,

$$SS3 = \frac{1}{NS}$$

where $SS3$ = location score for a sentence

NS = number of sentences in the document

Query-Bias Method

- Bias factor to score sentences containing query terms more highly.

The query-bias score $SS4$ is computed,

$$SS4 = \frac{tq^2}{nq}$$

where tq = number of query terms present in a sentence

nq = number of terms in a query

As in $SS2$, nq again acts as a normalisation factor for the query bias factor.

Combining the Scores

- The final score for each sentence is calculated by summing the individual score factors obtained for each method used.

$$SSS = aSS1 + bSS2 + cSS3 + dSS4$$

where SSS = sentence significance score

a, b, c, d are experimentally determined
scalar constants to control the influence of each factor

- The optimal length of a summary is a compromise between:
  - material covered in the summary.
  - appropriate length of summary, e.g. space available to display, time available to read.

Combining the Scores

- $SS1$, $SS2$, $SS3$ and $SS4$ need to be in similar number ranges to be able to balance their impact on the overall sentence score.
- The normalisation of $SS2$ and $SS4$, and selection of suitable values of $SS3$ value ensure that the combination process can be balanced by selection of suitable values of a, b, c and d.

Summarisation in Information Retrieval

This sentence-based summarisation method can be used to generate snippet summaries for inclusion in the SERP of an IR system.

Ideally these snippets enable the relevance of each document to the information need to be determined.

Snippets for IR strongly favour the presence of query terms in each sentence in the overall sentence score, which suggests giving this component a relatively high value in the above equation for SSS.

Summarisation in Information Retrieval

- In the context of the web, similar to search itself, the snippet generation process in web search needs to be robust enough to handle diverse forms of content structure over which the creator of the search engine has no control.
- The form of web content is determined by the author of the content or the designer of each webpage and website.
- How might this robustness in snippet creation be ensured?

Evaluation of Document Snippets in IR

How might we evaluation the effectiveness of a summarization process?

- Consider what recall and precision would mean in terms of the contents of a summary of a source document?

How could we evaluate the effectiveness of a snippet summary in a SERP?