Dublin City University

School of Computing

CA4009: Search Technologies

Section 4: Summarisation

Gareth Jones

October 2016

Introduction

Simply requiring the user to read a whole document looking for relevant information will often be very inefficient.

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

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

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

Introduction

Snippet creation is thus 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.

Document *summarisation* is a long established area of research.

- Similar to IR, summarisation was one of the earliest application areas explored for computers.

In this section we will explore the general concepts of summarization, 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 can have a complicated relationship with its source text.

A summary can be viewed as an open-vocabulary description of a document. The summary usually has a higher density of words describing the subject of the document than the original document.

Factors Affecting Summarisation

In general, 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.

Factors Affecting Summarisation

In addtion, to the issue of determining what the user needs from the summary, how to make a summary will depend on a number of factors.

- The form of the source text.

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

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

Factors Affecting Summarisation

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?

Factors Affecting Summarisation

The output will also be strongly influenced by multiple factors:

- the input form, e.g. natural language text, bullet points, tables, etc.
- the purpose of the summary.

Since the forms and needs of summaries are so varied, it seems unlikely that we can to expect to develop a single general technique for automatic summarization.

Instead we should look to develop a range of summarization techniques which are suitable for different situations, although even following this strategy is challenging.

Human Summarising

Summaries are traditionally created from source texts by humans, 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.

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

Automatic Summarisation Methods

Automatic summarisation can be broadly divided into two classes:

- **Information Extraction and Synthesis**: Information is extracted from the document using natural language processing methods and a new text synthesised using automatic text generation methods.
  - The summary contains information extracted from the original text, but is itself a new document.
  - 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.
  - Tex synthesis: text structure planning; selection of information from database; natural language text generation.

  We will not consider this approach further in this module.

Automatic Summarisation Methods

- **Sentence and/or Phrase Extraction**: Summary is composed of a subset of the sentences and/or phrases taken directly from the original document.
  - 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 one as the summary.
    Perhaps consider context in deciding which ones to use, e.g. is the sentence likely to make sentence without the one before it, if it starts with a pronoun.

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:

- Sentences may be ranked using some metric (or more likely a combination of metrics).
- Top n scoring sentences can then be taken, or sufficient sentences taken to reduce the document to $m\%$ of its original length.

These summaries can be difficult to read, 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

- 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 probably relationship to each other.
- Also as we noted earlier, *significant* words occur between low and high frequency limits.
- Luhn determined that two significant words can be considered to be 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 each 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.

Summarisation in Information Retrieval

This method can be used to generate snippet summaries for inclusion in the results page 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.

Summarisation in Information Retrieval

The search engine results page (SERP) of a web search engine often includes:

- selected images and/or
- a summary of the top level structure of returned websites. which it determines may form suitable relevant items in their own right. Providing this information potentially helps the user to determine relevance or navigate to the most useful part of the website.

Creation of these summaries is not covered formally in this module, but, based on what we do cover in the module. you should be able to suggest methods which could be used to create them.

Evaluation of Document Snippets in IR

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

Consider what recall and precision would mean for a summary of a source document?