Automatic summarizer – manual

Introduction

The Automatic Summarizer creates a summary by assigning values to each sentence based on specific statistical features of the text. After combining for each sentence the different feature values, the highest scoring sentences are displayed. This summarizer is *specifically tailored for scientific papers* and therefore will not perform very good on regular texts.

Download the jar file consisting of the source code and class files (http://martijnwieling.nl/files/SummarizerComplete.jar) and start the summarizer (using J2SE RE, version ≥ 1.4.2) via: `java -jar SummarizerComplete.jar`

A detailed report about the summarizer can be found here:
http://martijnwieling.nl/files/wielingvisser05automaticsummarization.pdf

And the associated powerpoint can be found here:
http://martijnwieling.nl/files/nlppres.ppt

Sentence features

Five features are used to calculate the score of a sentence:

- **Cue feature:** The occurrence of certain phrases in the document influences the cue value of the sentence. The phrases are specifically tailored for scientific papers (the files with the phrases used are included in the jar-archive). The score of a sentence can be negative (e.g. by including a phrase like 'for example'), or positive (i.e. by including a phrase like 'concluding' or 'the results').
- **Key feature:** The frequency of each uncommon word is counted and the sentence gets a score based on the number of high-frequent words (common words like 'that' or 'and' are filtered out). Two methods are available for calculating the key value: The first is Edmundson's key method, which assigns a score to each word equal to it's frequency. The key score of the sentence is calculated by summing the scores of it's words. The second is Luhn's key method, which doesn't use the exact frequency of each word, but treats each significant word in the same way. In contrast to Edmundson's method, it also takes the relative position of significant words into account. In Luhn's method a maximum number of non significant words (n.s. words) is specified. Within a sentence a range of words is selected which has a significant word at the beginning and end and doesn't have more non-significant words in this range than is specified. The value of a sentence is then calculated by taking the square of the number of significant words and dividing them by the total number of words in the range, e.g. for the sentence:
  * - - - [* - * * - - * - - *] - - * (with *: significant word, -: non-significant word, [...]: range, max n.s.: 5) the score is 5*5 / 10 = 2.5.

- **Title feature:** Each word in the sentence gets a score based on it's occurrence in the title or in one of it's headings. The title value of the sentence is calculated by summing the scores of it's words.
- **Location feature:** Sentences get a score based on their location (paragraph initial, paragraph final, in the first paragraph or in the last paragraph) and their occurrence beneath certain headings (like 'conclusions' or 'introduction).
- **Query feature:** The query score of a sentence is calculated by matching the sentence to the user specified query.

Using the Automatic Summarizer

Formatting the source text with Tag-buttons

The first step to generate a summary is to paste the source text in the large text area. If the text only consists of a title and headings of the same level, no tags have to be added. In that case the title is the first line of the source text, after which a blank line must be inserted. Every heading should be preceded and followed by a blank line. In the case of a paragraph (non-heading) of a single sentence, this sentence *must* be tagged by the **<T>** (Text) tag, even if this sentence spans multiple lines. Tagging can be done by highlighting the sentence(s) you wish to tag and pressing the corresponding tag button below the source text.

If the source text contains multi-level headings (e.g. headings of sections and headings of subsections), these can be tagged as follows. Do not tag the headings, but tag the subheadings with the tags starting from **<H2>** for each sublevel. Please also make sure that single line paragraphs are tagged by the **<T>** tag.

It is also possible to tag the text completely. This should be done in the following way. The title must be tagged with the **<H0>** (Title) tag. The highest level headings after the title (e.g. headings of sections) must be tagged with the **<H1>** (Heading 1) tag. Lower headings (e.g. headings of subsections) should be tagged with the following **<H#>** tags. Note that headings of the same level should be tagged with the same tag. Finally the complete text below each lowest-level heading should be selected and tagged with the **<T>** (Text) tag. Only one text-block can appear below each lowest level heading (so the text-block can be multiple paragraphs long, must always contain all text below the heading and may never contain other headings). To remove tags from the (part of the) source text, select the desired part and press the **Delete Tags** button.

To make sure certain sentences appear in the summary (e.g. author information), these sentences should be highlighted and be tagged with the **<F>** (Forward to summary) tag. Note that the number of sentences which can be forwarded is limited by the size of the summary.

A parser is used to separate the source text into sentences. In most cases this will succeed, however it is possible that a single sentence gets split into two other (incorrect) sentences (e.g. because of initials, or abbreviations not in the text-file). In this case the cursor should be placed after the token which forces the sentence split (e.g. a point (.)) and the **<NS>** button should be pressed. In this case the sentence will not be split. It is also possible that two sentences are incorrectly merged into one

sentenc. In this case the cursor should be placed after the first sentence and the **<S>** button should be split. In this case the sentences will not be merged.

Parameters and buttons

- **Source Language:** This should be set to the language of the source text, either Dutch or English.
- **Summary Size:** This defines the size of the summary as a percentage of the number of words of the source text (real value), e.g. if the source text contains 1000 words and the summary size is 25% the summary will contain approximately 250 words (excluding heading and title words). **The number of words of the source text** can be obtained by reading its tooltip (hover the mouse over the textarea to read it).
- **(Don't) Include Headings**: This defines if the summary contains the selected sentences and their headings, or only the selected sentences.
- **Query**: Here a user specified query can be entered. The query is NOT case sensitive. This field can be empty. Sentences which match the query are more likely to appear in the summary. The syntax of the query-field is as follows: each word which is entered must appear in the sentence, unless the OR keyword is placed between them. Note that nesting is not possible, so (keyword1) OR (keyword2) means: "(keyword1)" OR "(keyword2)". Common words, like 'and', are normally ignored, unless the '+' is typed before them. To specify a word which must not appear in the sentence, the sign '-' should precede the word. Exact phrase matches can be specified by quoting the phrase. An example of a query: `keyword1 OR +and keyword2 OR keyword2 -illegalWord OR "this is a phrase"`.
- **[New Summary]**: Press this button to clear the query textfield, the source textarea and the summary textarea.
- **[Reset Parameters]**: Press this button to reset all parameters to their default values.
- **Feature Weights**: The Integer values in these five textfields define the relative weights of each feature (as described above) to calculate the total weight of a sentence. If a feature should be ignored, a 0 should be entered in the corresponding textfield. It is also possible to enter negative values.
- **Key Method**: This is the key method which is used to calculate the key value of a sentence (the methods are described above, '# n.s. words' means 'a maximum of # non significant words').
- **Key Method, Lower Frequency Threshold (%)**: This real value defines what the least frequency of a word is (as a percentage of the total number of words in the source text) to mark it as significant (Luhn key method) or get a non-zero key score (Edmundson). So if the total number of words of the source text is 2500 and 0.2% is used as the threshold, only words which occur at least 5 times are used as significant words to calculate the key score of a sentence. To get an approximate value of the **absolute frequency threshold**, read the tooltip of this textfield.
- **Location Method Sentence Weights**: The Integer values in these four textfields define the relative weights which are used for the location feature (see above).
- **Title Method Weights**: The Integer values in these two textfields define the relative weights which are used for the title feature (see above). The first field is the score given to a word which matches a word in the title, the second field is the score given to a word which matches a word in one of it's headings.

- **[Summarize Text]**: Press this button to summarize the text, based on the entered parameters, the possible empty query and the source text. The result can be copied and stored on your system.

License

Authors

The automatic summarizer was designed and implemented by Martijn Wieling and Wicher Visser, graduated MSc students from the University of Groningen, Department of Computer Science, Intelligent Systems (2007).

Credits

We would like to thank:

- Simone Teufel and Marc Moens, for making their English cue phrase list and stop-list available to us
- Roeland Ordelman, to make a Dutch abbreviation list available to us
- Gertjan van Noord, for his support during this project