Processing of large document collections

Part 5 (Text summarization)

Helena Ahonen-Myka

Spring 2006

In this part

- text summarization, surface level methods
  - Luhn's method
  - Edmundson's method
  - corpus-based approaches: KPC method

2

Classical approaches

- Luhn '58
- general idea:
  - give a score to each sentence
  - choose the sentences with the highest score to be included in the summary

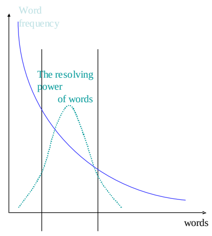
3

Luhn's method

- for each document:
  - filter terms in the document using a list of stopwords
  - normalize terms by stemming
    - differentiate, different, differently, difference -> differen
  - calculate frequencies of normalized terms
  - remove non-frequent terms
  - > "significant" terms remain

4

Luhn's method



- significant words occur "somewhat" frequently (within a document)
  - 2 thresholds (or a list of stopwords and 1 threshold)
- important sentences contain significant words
  - a sentence gets a higher score, if it contains more significant words

5

Luhn's method

- sentences are weighted using the resulting set of "significant" terms and a term density measure:
  - each sentence is divided into segments bracketed by significant terms not more than 4 non-significant terms apart
  - each segment is scored by taking the square of the number of bracketed significant terms divided by the total number of bracketed terms
    - score(segment) = $\text{significant\_terms}^2/\text{all\_terms}$

6

Exercise (CNN News)

- Let {13, computer, servers, Internet, traffic, attack, officials, said} be significant terms.
- ”Nine of the 13 computer servers that manage global Internet traffic were crippled by a powerful electronic attack this week, officials said.”

7

Exercise (CNN News)

- Let {13, computer, servers, Internet, traffic, attack, officials, said} be significant terms.
- * * * [13 computer servers * * * Internet traffic] * * * * * * [attack * * officials said]

8

Exercise (CNN News)

- [13 computer servers * * * Internet traffic]
  - score: 5^2 / 8 = 25/8 = 3.1
- [attack * * officials said]
  - score: 3^2 / 5 = 9/5 = 1.8

9

Luhn’s method

- the score of the highest scoring segment is taken as the sentence score
- the highest scoring sentences are chosen to the summary
- a cutoff value is given, e.g.
  - N best terms, or
  - x% of the original text

10

”Modern” application

- text summarization of web pages on handheld devices (Buyukkokten, Garcia-Molina, Paepcke; 2001)
- macro-level summarization
- micro-level summarization

11

Web page summarization

- macro-level summarization of a web page
  - the page is partitioned into ‘Semantic Textual Units’ (STUs)
    - paragraphs, lists, alt texts (for images)
  - hierarchy of STUs is identified
    - list - list item, table – table row
  - nested STUs are hidden

12

Web page summarization

- micro-level summarization: 5 methods tested for displaying STUs in several states
  - incremental: 1) the first line, 2) the first three lines, 3) the whole STU
  - all: the whole STU in a single state
  - keywords: 1) important keywords, 2) the first three lines, 3) the whole STU

13

Web page summarization

- summary: 1) the STUs 'most significant' sentence is displayed, 2) the whole STU
- keyword/summary: 1) keywords, 2) the STUs 'most significant' sentence, 3) the whole STU

- the combination of keywords and a summary has given the best performance for discovery tasks on web pages

14

Web page summarization

- extracting summary sentences
  - sentences are scored using a variant of Luhn's method:
    - words are TF*IDF weighted; given a weight cutoff value, the high scoring words are selected to be significant terms
    - weight of a segment: sum of the weights of significant words divided by the total number of words within a segment

15

Edmundson's method

- Edmundson (1969): New methods in automatic extracting
- extends earlier work to look at three features in addition to word frequencies:
  - cue phrases (e.g. "significant", "impossible", "hardly")
  - title and heading words
  - location

16

Features

- Location. Weight assigned to a text unit based on whether it occurs in lead, medial, or final position in a paragraph or the entire document, or whether it occurs in prominent sections such as the document's intro or conclusion
- Cue. Weight assigned to a text unit in case lexical or phrasal in-text summary cues occur: positive weights for bonus words (*"significant"*, *"verified"*, etc.), negative weights for stigma words (*"hardly"*, *"impossible"*, etc.)
- Key. Weight assigned to a text unit due to the presence of statistically significant terms (e.g., *tf or tf.idf* terms) in that unit
- Title. Weight assigned to a text unit for terms in it that are also present in the title, headline, initial paragraph (or the user's profile or query)

17

Combining the features

- methods to weight sentences based on each of the four features
  - weight of a sentence = the sum of the weights for features

$$Weight(U) = \alpha * Location(U) + \beta * Cue(U) + \\ + \chi * Key(U) + \delta * Title(U)$$

- *U* is a text unit such as a sentence, *Greek letters* denote weights of features

18

Evaluation

- methods were evaluated by comparison against manually created extracts
- corpus-based methodology: training and test sets
  - in the training phase, weights of the features were manually readjusted
- results:
  - three additional features dominated word frequency measures
  - the combination of cue-title-location was the best, with location being the best individual feature
  - keywords alone was the worst

19

Corpus-based approaches

- in the classical methods (Luhn, Edmundson), various features (thematic features, title, location, cue phrase) were used to determine the importance of information for summarization
- an obvious issue: determine the relative contribution of different features (tuning parameters) to any given text summarization task

20

Corpus-based approaches

- contribution of each feature is dependent on the text genre, e.g. location:
  - in newspaper stories, the leading text often contains a summary
  - in TV news, a preview segment may contain a summary of the news to come
  - in scientific text: an author-written abstract

21

Corpus-based approaches

- the importance of different text features for any given summarization problem can be determined by counting the occurrences of such features in text corpora
- in particular, analysis of human-generated summaries, along with their full-text sources, can be used to learn rules for summarization

22

Corpus-based approaches

- challenges
  - creating a suitable text corpus
  - ensuring that a suitable set of summaries is available
    - may already be available: scientific papers
    - if not: author, professional abstractor, judge
  - evaluation in terms of accuracy on unseen test data
  - discovering new features for new genres

23

KPC method

- Kupiec, Pedersen, Chen (1995): A trainable document summarizer
- a learning method using
  - a corpus of journal articles and
  - abstracts written by professional human abstractors
- naïve Bayesian classification method is used to create extracts

24

KPC method: general idea

- training phase:
  - select a set of features
  - calculate a probability of each feature value to appear in a summary sentence
    - using a training corpus (e.g. originals + manual summaries)

25

KPC method: general idea

- when a new document is summarized:
  - for each sentence
    - find values for the features
    - calculate the probability for this feature value combination to appear in a summary sentence
  - choose N best scoring sentences

26

KPC method: features

- 5 types of features
  - sentence-length cut-off feature
  - paragraph feature
  - thematic word feature
  - fixed-phrase feature
  - uppercase word feature

27

KPC method: features

- sentence-length cut-off feature
  - given a threshold (e.g. 5 words), the feature is true for all sentences longer than the threshold, and false otherwise
    - F1(s) = 0, if sentence s has 5 or less words
    - F1(s) = 1, if sentence s has more than 5 words

28

KPC method: features

- paragraph feature
  - only sentences in the first 10 paragraphs and the last 5 paragraphs in a document are taken into account
  - in paragraphs: paragraph-initial, paragraph-final, paragraph-medial are distinguished
    - F2(s) = i, if sentence s is the first sentence in a paragraph
    - F2(s) = f, if there are at least 2 sentences in a paragraph, and s is the last one
    - F2(s) = m, if there are at least 3 sentences in a paragraph, and s is neither the first nor the last sentence

29

KPC method: features

- thematic word feature
  - a small number of thematic words (the most frequent content words) are selected
  - each sentence is scored as a function of frequency of the thematic words
  - highest scoring sentences are selected
  - binary feature: feature is true for a sentence, if the sentence is present in the set of highest scoring sentences

30

KPC method: features

- fixed-phrase feature
  - this feature is true for sentences
    - that contain any of 26 indicator phrases (e.g. ”this letter…”, ”In conclusion…”), or
    - that follow section head that contains specific keywords (e.g. ”results”, ”conclusion”)

31

KPC method: features

- uppercase word feature
  - proper names and explanatory text for acronyms are usually important
  - feature is computed like the thematic word feature (binary feature)
  - an uppercase thematic word is not sentence-initial and begins with a capital letter and must occur several times
  - first occurrence is scored twice as much as later occurrences

32

Exercise (CNN news)

- in our example, we use 3 (modified) features
- feature *sentence-length*; F1: let threshold = 14
  - < 14 words: F1(s) =0, else F1(s)=1
- feature *paragraph*; F2:
  - i=first, f=last, m=medial
- feature *thematic-words*; F3
  - score: how many thematic words a sentence has
  - F3(s) = 0, if score > 3,
    else F3(s) = 1

33

KPC method: classifier

- for each sentence s, we compute the probability that s will be included in a summary S given the k features F_j, j=1...k
- the probability can be expressed using Bayes' rule:

$$P(s \in S \mid F_1, \ldots, F_k) = \frac{P(F_1, \ldots, F_k \mid s \in S) P(s \in S)}{P(F_1, \ldots, F_k)}$$

34

KPC method: classifier

- assuming statistical independence of the features:

$$P(s \in S \mid F_1, \ldots F_k) = \frac{\prod_{j=1}^{k} P(F_j \mid s \in S) P(s \in S)}{\prod_{j=1}^{k} P(F_j)}$$

- *P(s∈S)* is a constant, and *P(F_j| s∈S)* and *P(F_j)* can be estimated directly from the training set by counting occurrences

35

KPC method: corpus

- corpus was acquired from a company which provides abstracts of technical articles to online information services
- articles did not have author-written abstracts
- abstracts were created by professional abstractors

36

KPC method: corpus

- 188 document/summary pairs sampled from 21 publications in the scientific/technical domain
- summaries were mainly indicative, average length was 3 sentences
- average number of sentences in the original documents was 86
- author, address, and bibliography were removed

37

KPC method: sentence matching

- the abstracts from the human abstractors were not extracts but inspired by the original sentences
- the automatic summarization task here:
  – extract sentences that the human abstractor might have chosen to prepare summary text (with minor modifications...)
- for training, a correspondence between the manual summary sentences and sentences in the original document needed to be obtained

38

KPC method: sentence matching

- matching can be done in several ways:
  – a direct sentence match
    - the same sentence is found in both
  – a direct join
    - 2 or more original sentences were used to form a summary sentence
  – summary sentence can be 'unmatchable'
  – summary sentence (single or joined) can be 'incomplete'

39

KPC method: sentence matching

- matching was done in two passes
  – first, the best one-to-one sentence matches were found automatically
  – second, these matches were used as a starting point for the manual assignment of correspondences

40

KPC method: evaluation

- cross-validation strategy for evaluation
  – documents from a given journal were selected for testing one at a time
  – all other document/summary pairs (of this journal) were used for training
  – results were summed over journals
- unmatchable and incomplete summary sentences were excluded
- total of 498 unique sentences

41

KPC method: evaluation

- two ways of evaluation
  – the fraction of manual summary sentences that were faithfully reproduced by the summarizer program
    - the summarizer produced the same number of sentences as were in the corresponding manual summary
    - -> 35% of summary sentences reproduced
    - 83% is the highest possible value, since unmatchable and incomplete sentences were excluded
  – the fraction of the matchable sentences that were correctly identified by the summarizer
    - -> 42%

42

KPC method: evaluation

- the effect of different features was also studied
  - best combination (44%): paragraph, fixed-phrase, sentence-length
  - baseline: selecting sentences from the beginning of the document (result: 24%)
- if 25% of the original sentences selected: 84%
- conclusion: comparable to manually tuned feature weights (or better)

43

Text summarization

- next time:
  - discourse-based text summarization
  - multi-document summarization
  - summarizing database content

44