Processing of large document collections

Part 5 (Text summarization)
Helena Ahonen-Myka
Spring 2005

In this part

- text summarization
  - Edmundson's method
  - corpus-based approaches: KPC method

2

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

3

Edmundson's method

- programs to weight sentences based on each of the four features
  - weight of a sentence = the sum of the weights for features
- programs were evaluated by comparison against manually created extracts
- corpus-based methodology: training set and test set
  - in the training phase, weights were manually readjusted

4

Edmundson's method

- 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

5

Fundamental issues

- What are the most powerful but also more general features to exploit for summarization?
- How do we combine these features?
- How can we evaluate how well we are doing?

6

Linear Weighting Scheme

$$Weight(U) := \alpha * Location(U) + \beta * CuePhrase(U) + \chi * StatTerm(U) + \delta * AddTerm(U)$$

U is a text unit such as a sentence, *Greek letters* denote tuning parameters

- 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
- CuePhrase 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.)
- StatTerm Weight assigned to a text unit due to the presence of statistically salient terms (e.g., *tf.idf* terms) in that unit
- AddTerm Weigth assigned to a text unit for terms in it that are also present in the title, headline, initial para, or the user's profile or query

7

Corpus-based approaches

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

8

Corpus-based approaches

- contribution 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

9

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

10

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

11

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 (Engineering Information Co.)
- naïve Bayesian classification method is used to create extracts

12

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)

13

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

14

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

15

KPC method: features

- paragraph feature
  - sentences in the first 10 paragraphs and the last 5 paragraphs in a document get a higher value
  - 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

16

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

17

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 contain specific keywords (e.g. "results", "conclusion")

18

KPC method: features

- uppercase word feature
  - proper names and explanatory text for acronyms are usually important
  - feature is computed like the thematic word 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

19

Exercise (CNN news)

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

20

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:

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

21

KPC method: classifier

- assuming statistical independence of the features:

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

- $P(s \in S)$ is a constant, and $P(F_j \mid s \in S)$ and $P(F_j)$ can be estimated directly from the training set by counting occurrences

22

KPC method: corpus

- corpus is acquired from Engineering Information Co, which provides abstracts of technical articles to online information services
- articles do not have author-written abstracts
- abstracts were created by professional abstractors

23

KPC method: corpus

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

24

KPC method: sentence matching

- the abstracts from the human abstractors are 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 need to be obtained
- matching can be done in several ways

25

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'

26

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

27

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

28

KPC method: evaluation

- two ways of evaluation
  1. 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
  2. the fraction of the matchable sentences that were correctly identified by the summarizer
     - -> 42%

29

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%

30