Processing of large document collections

Part 7 (Text summarization: multi-document summarization, knowledge-rich approaches, current topics)

Helena Ahonen-Myka

Spring 2005

In this part…

- Summarization of multiple documents
  - MEAD
- Knowledge-rich approaches
  - STREAK
- Current topics in text summarization

Summarization of multiple documents

- Radev, et al (2004): Centroid-based summarization of multiple documents
- idea: summarizing news events
  - news stories come from several sources (e.g. news agencies)
  - all news stories talking about the same event (e.g. accident, earthquake,…) are clustered
    - stories in one cluster repeat (partially) the same content
    - stories have a chronological order (time stamp)
  - one summary for each cluster is created
    - a reader does not have to read the same content several times

Centroid-based clustering

- each document is a tf * idf weighted vector
- documents are clustered
  1. cluster centroid = first document
  2. a new document D is compared to each centroid C
     1. if sim(C, D) > threshold, D is included in C, and C is updated
     2. if D is not included in any cluster, it becomes the centroid of a new cluster

$$S(C_i, D_j) = \frac{\sum_{k=1}^{|T|} w_{ki} \cdot w_{kj} \cdot idf(k)}{\sqrt{\sum_{k=1}^{|T|} w_{ki}^2} \cdot \sqrt{\sum_{k=1}^{|T|} w_{kj}^2}}$$

MEAD extraction algorithm

- sentences are ranked according to a set of features
- input:
  - a cluster of documents, segmented into n sentences
  - compression rate r
- output:
  - a sequence of n x r sentences from the original documents
    - presented in the same order as in the input documents

Features

- three features:
  - centroid value C_i for sentence S_i is the sum of the centroid values of all words in the sentence

$$C_i = \sum_{w \in S_i} C_w$$

- the centroid vector of the cluster represents importance of words for all the documents in the cluster

Features

- positional value:
  - C_{max} = score of the highest-ranking sentence in the document according to the centroid value
  - the i^{th} sentence in a document gets a value

$$P_i = \frac{(n-i+1)}{n} \times C_{\max}$$

7

Features

- first sentence overlap F_i:
  - the inner product of the current sentence S_i and the first sentence of the document
- combined score of sentence S_i: linear combination of three features
  - score(S_i) = $w_c C_i + w_p P_i + w_f F_i$

8

Cross-sentence dependencies

- scores of sentences can be further refined after considering possible cross-sentence dependencies, for instance
  - repeated content in sentences
    - redundant content can be removed
  - chronological ordering
    - earlier or later sentences can be preferred
  - source preferences
    - e.g. Helsingin sanomat is trusted more than Iltalehti…

9

Repeated content

1. John Doe was found guilty of the murder.
2. The court found John Doe guilty of the murder of Jane Doe last August and sentenced him to life. (2. presents additional content -> 1. redundant)
3. Eighteen decapitated bodies have been found in a mass grave in northern Algeria, press reports said Thursday.
4. Algerian newspapers have reported on Thursday that 18 decapitulated bodies have been found by the authorities. (equivalent content)

10

Reranking based on repeated content

- redundancy penalty R_{ij} for each sentence i which overlaps with sentences j that have higher score value

$$R_{ij} = 2 \times \frac{\#overlapping\ words}{\#words\ in\ sentence_i + \#words\ in\ sentence_j}$$

- redundancy penalty for a sentence: max (R_{ij})
- new_score(s_i) = $w_c C_i + w_p P_i + w_f F_i - w_R R_i$
- all sentences are reranked by new_score and a new extract in created
- iteration until reranking does not result in a different extract

11

Knowledge-rich approaches

- structured information can be used as the starting point for summarization
  - structured information: e.g. data and knowledge bases
  - may have been produced by processing input text (information extraction)
- summarizer does not have to address the linguistic complexities and variability of the input, but also the structure of the input text is not available

12

Knowledge-rich approaches

- there is a need for measures of salience and relevance that are dependent on the knowledge source
- addressing cohesion, and fluency becomes the entire responsibility of the generator

13

STREAK

- McKeown, Robin, Kukich (1995): Generating concise natural language summaries
- goal: folding information from multiple facts into a single sentence using concise linguistic constructions

14

STREAK

- produces summaries of basketball games
- first creates a draft of essential facts
- then uses revision rules constrained by the draft wording to add in additional facts as the text allows
  - revision rules have been extracted by studying human-written game summaries

15

STREAK

- input:
  - a set of box scores for a basketball game
  - historical information (from a database)
- task:
  - summarize the highlights of the game, underscoring their significance in the light of previous games
- output:
  - a short summary: a few sentences

16

STREAK

- the box score input is represented as a conceptual network that expresses relations between what were the columns and rows of the table
- essential facts: the game result, its location, date and at least one final game statistic (the most remarkable statistic of a winning team player)

17

STREAK

- essential facts can be obtained directly from the box-score
- in addition, other potential facts
  - other notable game statistics of individual players - from box-score
  - game result streaks (Utah recorded its fourth straight win) - historical
  - extremum performances such as maximums or minimums - historical

18

STREAK

- essential facts are always included
- potential facts are included if there is space
  - decision on the potential facts to be included could be based on the possibility to combine the facts to the essential information in cohesive and stylistically successful ways

19

STREAK

- given facts:
  - Karl Malone scored 39 points.
  - Karl Malone's 39 point performance is equal to his season high
- a single sentence is produced:
  - Karl Malone tied his season high with 39 points

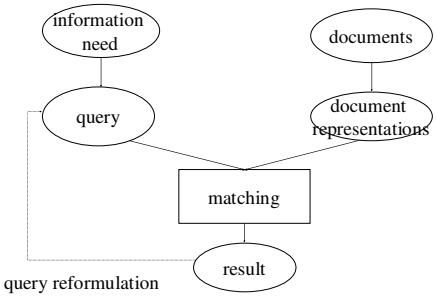
20

Current topics in text summarization

- multi-document summarization
- non-extrative summarization (abstracts)
- spoken language (incl. dialogue) summarization
- multilingual summarization
- integrating of question answering and text summarization
- web-based & multimedia summarization
- evaluation of summarization systems
  - Document Understanding Conferences (DUC)

21

Mapping to the information retrieval process



22