Processing of large document collections

Part 6 (Text summarization: discourse-based approaches)

Helena Ahonen-Myka

Spring 2005

Discourse-based approaches

- discourse structure appears to play an important role in the strategies used by human abstractors and in the structure of their abstracts
- an abstract is not just a collection of sentences, but it has an internal structure
  - --> abstract should be coherent and it should represent some of the argumentation used in the source

Boguraev, Kennedy (BG)

- Boguraev, Kennedy (1999): Salience-based content characterisation of text documents
- goal: identify those phrasal units across the entire span of the document that best function as representative highlights of the document's content
- these phrasal units are called topic stamps
- a set of topic stamps is called capsule overview

BG

- a capsule overview
  - not a set/sequence of sentences
  - a semi-formal (normalised) representation of the document, derived after a process of data reduction over the original text
  - not always very readable, but still represents the flow of the narrative
    - can be combined with surrounding information to produce more coherent presentation

Priest is charged with Pope attack

A Spanish priest was charged here today with attempting to murder the Pope. Juan Fernandez Krohn, aged 32, was arrested after a man armed with a bayonet approached the Pope while he was saying prayers at Fatima on Wednesday night.

According to the police, Fernandez told the investigators today that he trained for the past six months for the assault. He was alleged to have claimed the Pope 'looked furious' on hearing the priest's criticism of his handling of the church's affairs. If found quilty, the Spaniard faces a prison sentence of 15-20 years.

Capsule overview vs. summary

- summary could be, e.g.
  - "A Spanish priest is charged after an unsuccessful murder attempt on the Pope"
- capsule overview:
  - A SPANISH PRIEST was charged
  - Attempting to murder the POPE
  - HE trained for the assault
  - POPE furious on hearing PRIEST'S criticisms

BG

- primary consideration: methods should apply to any document type and source (domain independence)
- also: efficient and scalable technology
  - shallow syntactic analysis, no comprehensive parsing engine needed

7

BG

- based on the findings on technical terms
  - technical terms have such linguistic properties that can be used to find terms automatically in different domains quite reliably
  - technical terms seem to be topical
- task of content characterization
  - identifying phrasal units that have
    - lexico-syntactic properties similar to technical terms
    - discourse properties that signify their status as most prominent

8

Terms as content indicators

- problems
  - undergeneration
  - overgeneration
  - differentiation

9

Undergeneration

- a set of phrases should contain an exhaustive description of all the entities that are discussed in the text
- the set of technical terms has to be extended to include also expressions with pronouns etc.

10

Overgeneration

- already the set of technical terms can be large
- extensions make the information overload even worse
- solution: phrases that refer to one participant in the discourse are combined with referential links

11

Differentiation

- the same list of terms may be used to describe two documents, even if they, e.g., focus on different subtopics
- it is necessary to differentiate term sets not only according to their membership, but also according to the relative representativeness of the terms they contain

12

Term sets and coreference classes

- phrases are extracted using a phrasal grammar (e.g. a noun with modifiers)
  - also expressions with pronouns and incomplete expressions are extracted
  - using a (Lingsoft) tagger that provides information about the part of speech, number, gender, and grammatical function of tokens in a text
  - solves the undergeneration problem

13

Term sets and coreference classes

- the phrase set has to be reduced to solve the problem of overgeneration
- -> a smaller set of expressions that uniquely identify the objects referred to in the text
- application of anaphora resolution
  - e.g. to which noun a pronoun 'he' refers to?

14

Resolving coreferences

- procedure
  - moving through the text sentence by sentence and analysing the nominal expressions in each sentence from left to right
  - either an expression is identified as a new participant in the discourse, or it is taken to refer to a previously mentioned referent

15

Resolving coreferences

- coreference is determined by a 3 step procedure
  - a set of candidates is collected: all nominals within a local segment of discourse
  - some candidates are eliminated due to morphological mismatches or syntactical restrictions
  - remaining candidates are ranked according to their relative salience in the discourse

16

Salience factors

- sent(term) = 100 iff term is in the current sentence
- cntx(term) = 50 iff term is in the current discourse segment
- subj(term) = 80 iff term is a subject
- acc(term) = 50 iff term is a direct object
- dat(term) = 40 iff term is an indirect obj
- ...

17

Local salience of a candidate

- the local salience of a candidate is the sum of the values of the salience factors
- the most salient candidate is selected as the antecedent
- if the coreference link cannot be established to some other expression, the nominal is taken to introduce a new referent
- -> coreferent (equivalence) classes

18

Topic stamps

- in order to further reduce the referent set, some additional structure has to be imposed
  - the referent set is ranked according to the salience of its members (= coreference classes)
  - objects in the centre of discussion have a high degree of salience
  - salience is measured like local saliency in coreference resolution, but tries to measure the importance of unique referents in the discourse
    - salience of coreference class = sum of the salience factor values of the expressions in the class

19

Priest is charged with Pope attack

A Spanish priest was charged here today with attempting to murder the Pope. Juan Fernandez Krohn, aged 32, was arrested after a man armed with a bayonet approached the Pope while he was saying prayers at Fatima on Wednesday night.

According to the police, Fernandez told the investigators today that he trained for the past six months for the assault. He was alleged to have claimed the Pope 'looked furious' on hearing the priest's criticism of his handling of the church's affairs. If found quilty, the Spaniard faces a prison sentence of 15-20 years.

20

Saliency

- 'priest' is the primary element
  - eight references to the same actor in the body of the story
  - these references occur in important syntactic positions: 5 are subjects of main clauses, 2 are subjects of embedded clauses, 1 is a possessive
- 'Pope attack' is also important
  - 'Pope' occurs 5 times, but not in so important positions (2 are direct objects)

21

Capsule overview

- capsule overview:
  - occurrences of the topic stamps with some context
    - A SPANISH PRIEST was charged
    - Attempting to murder the POPE
    - HE trained for the assault
    - POPE furious on hearing PRIEST'S criticisms
- several granularities for context: phrase, sentence, paragraph…

22

Discourse segments

- if the intention is to use very concise descriptions of one or two salient phrases, i.e. topic stamps, longer text have to be broken down into smaller segments
- topically coherent, contiguous segments can be found by using a lexical similarity measure
  - assumption: distribution of words used changes when the topic changes

23

BG: Summarization process

1. linguistic analysis
2. discourse segmentation
3. extended phrase analysis
4. coreference resolution
5. calculation of discourse salience
6. topic stamp identification
7. capsule overview

24