Automatic Text Summarization

Martin Hassel
KTH CSC
School of Computer Science and Communication
xmartin@kth.se

Text Summarization

- To extract the gist, the essence, of a text and present it in a shorter form with as little loss as possible with respect to mediated information
- Redundancy (Shannon 1951)
  - Facilitates recovery in noisy channels

Automatic Text Summarization

- Automatic Text Summarization is the technique where a computer program summarizes a text
- The program is given a text and returns a shorter, hopefully non-redundant, text
- The earliest systems are from the 60's
- Luhn 1959, Edmunson 1969 and Salton 1989.

- The technique has been in development for more than 30 years
- Data storage was expensive - shortening of texts before indexing was needed
- New uses and interest in the area has arisen with the expansion of the Internet
- Today's computers are powerful enough to summarize large quantities of text quickly
- MS Word, Sherlock 2 (Mac OS)

Methods for Summarization

- Is done with linguistic as well as statistic and heuristic methods
- Abstraction vs. Extraction
- Single Document vs. Multi Document
- Minimal summary: keyword list, kwic

Text Abstraction

- Text abstraction – what humans do
- We read a text, reinterpret it, and rewrite it in our own words

- With a computer:
  - Semantic parsing
  - Translation into a formal language
  - A set of choices regarding what is to be said based on the formal description
  - Text generation (surface generation)
    - New syntactic structures
    - New lexical choices

Text Extraction

- Topic identification
- Statistic and heuristic methods
  - Keyword extraction
- Scoring
- Extract the most relevant/central text segments (i.e. paragraphs, sentences, phrases etc.) and concatenate them to form a new text
- Most automatic summarizers are extraction based

- Automatic Text Summarization is a far cry from human abstraction, and will probably never be as good
- BUT, it is faster and cheaper!

Methods for Text Extraction

Summarization methods and algorithms based on extraction (Chin-Yew Lin 1999):

- Baseline: Sentence order in text gives the importance of the sentences. First sentence highest ranking last sentence lowest ranking.
- Title: Words in title and in following sentences gives high score.

- Term frequency (tf): Open class terms which are frequent in the text are more important than the less frequent. Open class terms are words that change over time.
- Position score: The assumption is that certain genres put important sentences in fixed positions. For example: Newspaper articles has most important terms in the 4 first paragraphs.

- Query signature: The query of the user affect the summary in the way that the extract will contain these words (for example in a search engine).
- Sentence length: The sentence length implies which sentence is the most important.
- Average lexical connectivity: Number terms shared with other sentences. The assumption is that a sentence that share more terms with other sentences is more important.

- Numerical data: Sentences containing numerical data are scored higher than the ones without numerical values.
- Proper name: Dito for proper names in sentences.
- Pronoun and Adjective: Dito for pronouns and adjectives in sentences. Pronouns reflecting coreference connectivity.
- Weekdays and Months: Dito for Weekdays and Months:

Martin Hassel

- Quotation: Sentences containing quotations might be important for certain questions from user.
- First sentence: First sentence of each paragraphs are the most important sentences.
- Simple combination function: All the above parameters were normalized and put in a combination function with no special weighting.

Martin Hassel

Domain Terms

- *tf* = term frequency, number of unique terms (”words”) in a document
- *idf* = inverse document frequency, number of documents in which the term occurs divided with the total number of documents
- *tf*•*idf* measures how significant a term is for a document. Terms with a good *tf*•*idf* score are good descriptors of that document

Martin Hassel

SweSum

- Summarizes Swedish, English, Danish, Norwegian, French, German, Spanish, Italian, Persian and Greek newspaper text and shorter report texts online
- Formatting
  - HTML bold face
  - New paragraph
  - Headings
  - Titles

Martin Hassel

- User adaptation / slanting
  - User submitted keywords
- Naïve combination function
  - Utilizes aforementioned indicators
  - Each indicator is weighted
  - Each sentence is assigned a score

Martin Hassel

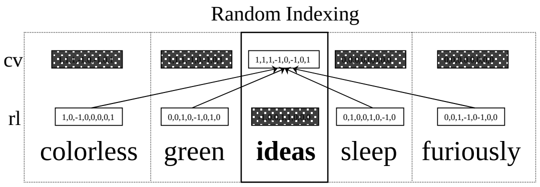
- For Swedish: lexicon with 700.000 open class words (conjugated form mapped to its lemma)
- 70-80% of central facts kept when keeping 30% of 3-4 pages of news text
- Implemented in Perl-CGI
- http://swesum.nada.kth.se/

Martin Hassel

HolSum

- Language independent summarizer
  - “small” languages lack large amounts of annotated or structured data
- Aims for overview summaries
  - try to find a summary of a given length as similar as possible to the original document

Capturing Context



cv = context vector
rl = random label

Capturing Content

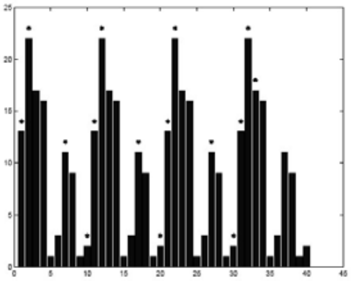
? How do we transform a document’s words’ conceptual representations into a content representation of the document

! By summing the *tf*•log(*idf*) weighted context vectors of the words that occur in the particular text

Finding a Better Summary

- Greedy search using initial summary:
  1) Transform summary candidate (remove / add one or two sentences)
  2) Compare new summary candidate to document
  3) Keep best candidate (old or new)
  4) Repeat 1-3 until no better summary is found
- Selecting summaries instead of sentences

Variation in Sentence Selection



Number of human produced extracts that included each sentence from one of the Swedish corpus texts. There is a total of 27 human produced extracts for this text. Sentences marked with a * are those selected by HolSum.

Challenging Topics

- Pronouns and other anaphoric phenomena
  - Pronoun resolution
- Sentences are often too large or too small to use as extraction units
  - Phrase reduction and combination rules

Pronoun Resolution

- Dangling anaphors
  - *Peter* ran. *He* ran as fast as he could.

With Pronouns Retained

Analysera mera!

Regi: Harold Ramis

Medv: Robert De Niro, Billy Crystal, Lisa Kudrow

Längd: 1 tim, 45 min

…

Ett av många skäl att glädjas åt Analysera mera är att Robert De Niro här verkligen utövar skådespelarkonst igen. Han accelererar emotionellt från 0 till 100 på ingen tid alls, för att sedan kattmjukt bromsa in och parkera, lugnt och behärskat. Och han är tämligen oemotståndlig. Här har han åstadkommit ännu en intelligent komedi för alla oss vänner av intelligens och komedi, gärna i kombination.

SvD 99-10-08

With Pronouns Resolved

Analysera mera!

Regi: Harold Ramis

Medv: Robert De Niro, Billy Crystal, Lisa Kudrow

Längd: 1 tim, 45 min

…

Ett av många skäl att glädjas åt Analysera mera är att Robert De Niro här verkligen utövar skådespelarkonst igen. **Robert** accelererar emotionellt från 0 till 100 på ingen tid alls, för att sedan kattmjukt bromsa in och parkera, lugnt och behärskat. Och **Robert** är tämligen oemotståndlig. Här har **Harold** åstadkommit ännu en intelligent komedi för alla oss vänner av intelligens och komedi, gärna i kombination.

SvD 99-10-08

Issues in Pronoun Resolution

- Nouns do not always indicate their gender
- Pronouns do not always refer linearly
- Identification of pronouns
  - Determiners
  - Cataphora

Pronoun Resolution in Practice

- Mitkov’s limited knowledge approach
  - Does not require parsing, only part-of-speech tagging and noun phrase chunking
  - More intuitive weighting system than Lappin & Leass
  - However, misses grammatical role cues
  - Successfully implemented for at least English, Polish and Arabic

Mitkov’s Algorithm

1. Take part-of-speech tagged text as input
2. Identify noun phrases at most 2 sentences away from the current anaphor
3. Check for number and gender agreement
4. Apply genre-specific antecedent indicators
5. Choose as antecedent the cantidate with highest indicator score

Mitkov’s Antecedent Indicators 1

- Definiteness
- Giveness
- Lexical reiteration
- Section heading preference
- Non-prepositional noun phrases
- Referential distance

Martin Hassel

Mitkov’s Antecedent Indicators 2

- Collocation pattern preference
- Immediate reference
- Genre specific indicators
  - Indicating verbs
  - Term preference

Martin Hassel

Mitkov’s Tie Breaking Scheme

- If two or more noun phrases share highest score, prefer the candidate:
  1. With the highest immediate reference score
  2. With the highest collocation pattern score
  3. With the highest indicating verb score
  4. Most recent of remaining candidates

Martin Hassel

Phrases as Smallest Extraction Unit

Phrase reduction and phrase combination rules (Hongyan Jing 2000):

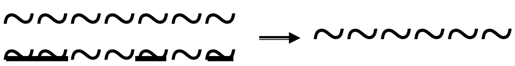
- The goals of reduction
  - remove as many redundant phrases as possible
  - do not detract from the main idea the sentence conveys
- The key problem
  - *decide when it is appropriate to remove a phrase*

Martin Hassel

Major Cut and Paste Operations

- (1) Sentence reduction

- (2) Sentence Combination

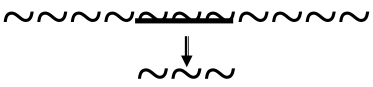


Martin Hassel

Major Cut and Paste Operations

- (3) Syntactic Transformation

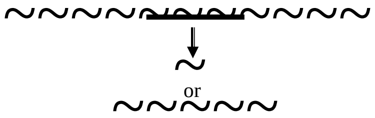
- (4) Lexical paraphrasing



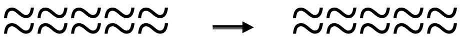
Martin Hassel

Major Cut and Paste Operations

- (5) Generalization/Specification



- (6) Sentence reordering



Sentence Reduction

Original Sentence: When it arrives sometime next year in new TV sets, the V-chip will give parents a new and potentially revolutionary device to block out programs they don’t want their children to see.

Reduction Program: The V-chip will give parents a new and potentially revolutionary device to block out programs they don’t want their children to see.

Professional: The V-chip will give parents a device to block out programs they don’t want their children to see.

Sentence Combination

S1: But it also raises serious questions about the privacy of such highly personal information *wafting about the digital world*.

S2: This issue thus fits squarely into the broader debate about privacy and security on the internet *whether it involves protecting credit card numbers or keeping children from offensive information*.

Combined: But it also raises serious questions about the privacy of such personal information and this issue thus fits squarely into the broader debate about privacy and security on the internet.

Applications

- Summaries of:
  - Newspaper text (for journalists, media surveillance, business intelligense etc).
  - Reports (for politicians, commissioners, businessmen etc).
  - E-mail correspondence
  - In search engines to extract key topics or to present summaries (instead of snippets) of the hits for easier relevance estimation

- Headline generation and minimal summaries for SMS on mobile phones
- Automatic compacting of web pages for WAP
- For letting a computer read summarized web pages by telephone (SiteSeeker Voice)
- To enable search in foreign languages and getting an automatic summary of the automatically translated text
- To facilitate identification of a specific document in a document collection

Text Summarizers

- Automated Text Summarization (SUMMARIST)
- Autonomy
- Intelligent Miner for Text - Summarization tool (IBM)
- Inxight (XEROX)
- Microsoft Word – AutoSummarize
- OracleContext
- Sherlock 2 (Mac OS).
- SweSum (KTH)