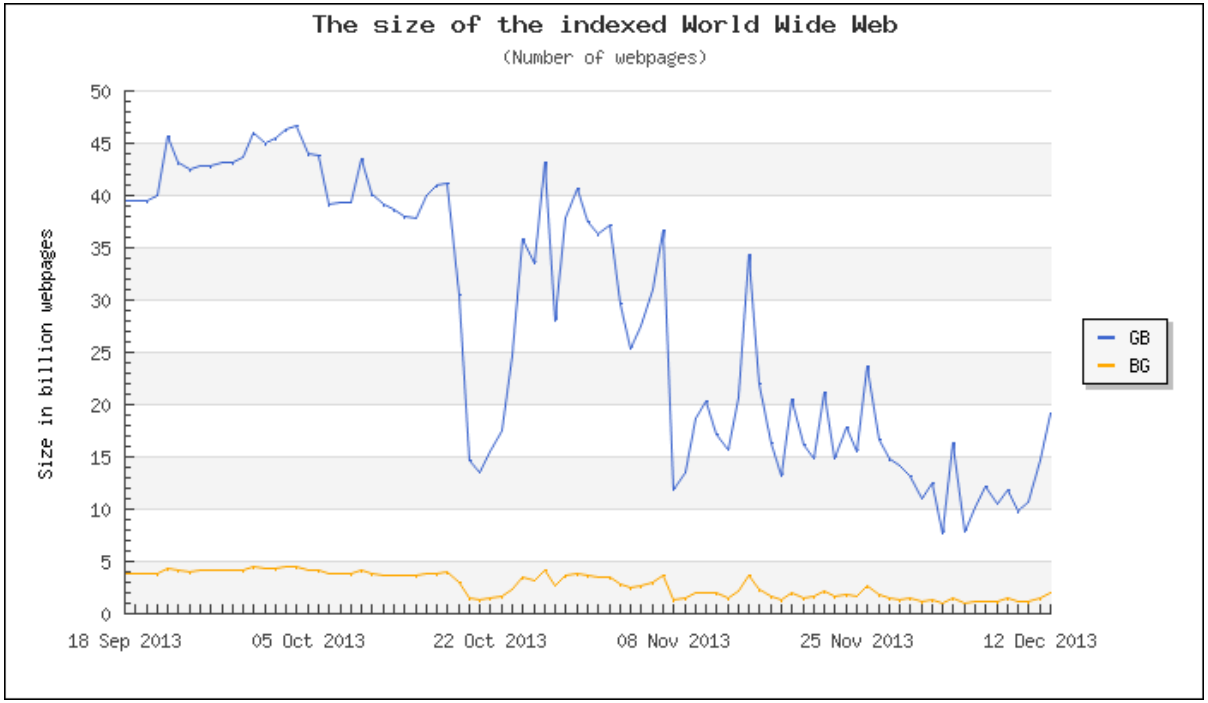
Multi-Document Summarization (MLTA 2014)

Dr. Tanveer J. Siddiqui

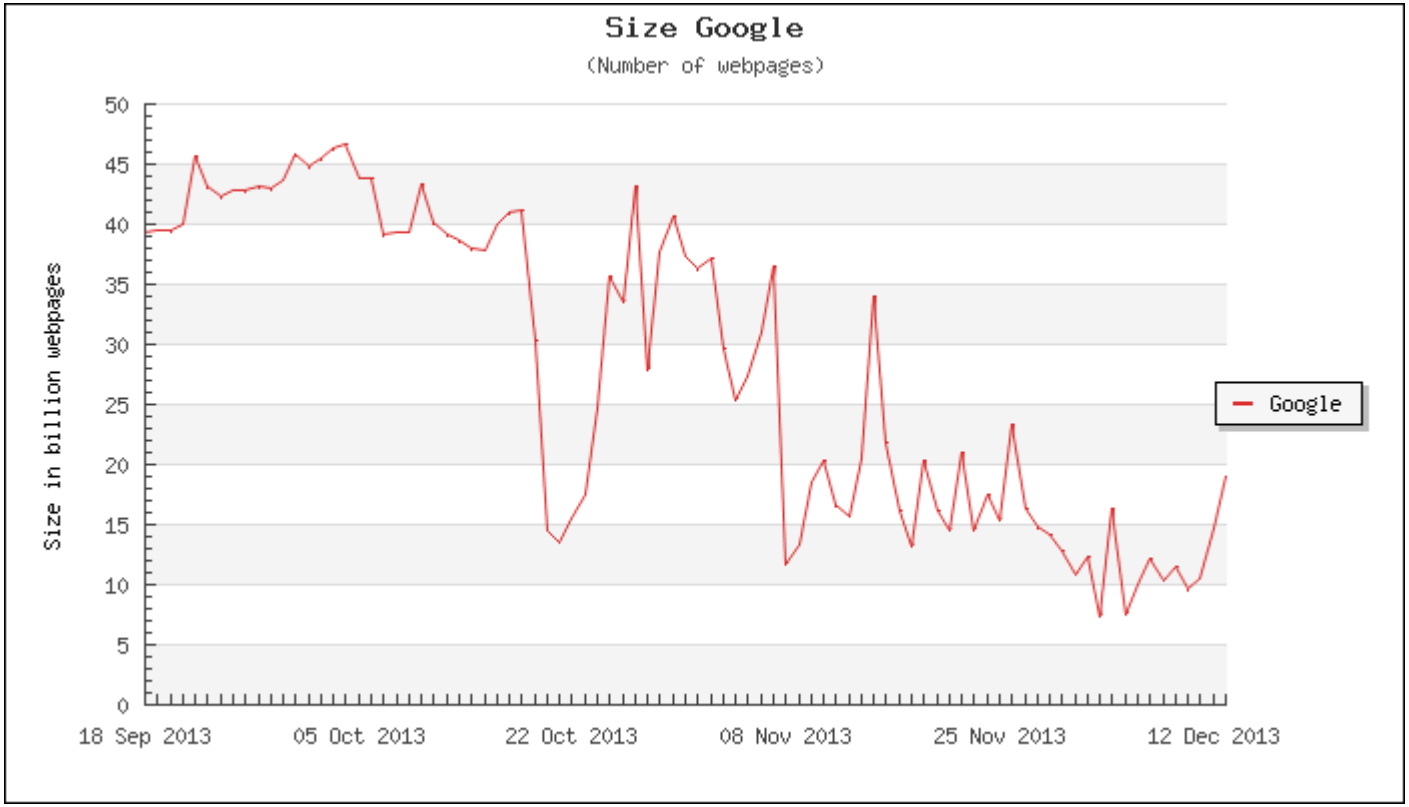
Outline

- Introduction to Text Summarization
- Some Real life examples
- Types of Summaries
- Early work
- Sentence Extraction Methods
- Evaluation
- Multi-document summarization

Information Overload



Source: http://www.worldwidewebsize.com/


Size Google
(Number of webpages)
Size in billion webpages
50
45
40
35
30
25
20
15
10
5
0
18 Sep 2013
05 Oct 2013
22 Oct 2013
08 Nov 2013
25 Nov 2013
12 Dec 2013
Google

Possible approaches

- information retrieval
- information extraction
- visualization
- question answering
- document clustering
- text summarization

Summarizer

- “A Summarizer is a system whose goal is to produce a condensed representation of the content of its input for human consumption” (Mani, 2001, p.3)

Summarization in Everyday Life

- News paper headline
- Preview or trailer of a show
- Abstract of a scientific articles
- Conference program
- Table showing baseball statistics
- Book reviews
- Weather forecast
- Library catalog
- Product list

Abstract of a technical paper

www.elsevier.com/locate/infoproman

Generic technologies for single- and multi-document summarization

Marie-Francine Moens *, Roxana Angheluta, Jos Dumortier

Interdisciplinary Centre for Law & IT (ICRI), Katholieke Universiteit Leuven, Tiensestraat 41, B-3000 Leuven, Belgium

Received 30 July 2003; accepted 17 December 2003
Available online 5 March 2004

Abstract

The technologies for single- and multi-document summarization that are described and evaluated in this article can be used on heterogeneous texts for different summarization tasks. They refer to the extraction of important sentences from the documents, compressing the sentences to their essential or relevant content, and detecting redundant content across sentences. The technologies are tested at the Document Understanding Conference, organized by the National Institute of Standards and Technology, USA in 2002 and 2003. The system obtained good to very good results in this competition. We tested our summarization system also on a variety of English Encyclopedia texts and on Dutch magazine articles. The results show that relying on generic linguistic resources and statistical techniques offer a basis for text summarization.

Movie trailer

Conference programme

Monday 8 December 2014

9:30 – 10:30: Registration & coffee

10:30 – 11:00: Welcome keynotes

Bernadette Dorizzi, Dean for Research, Telecom SudParis
Patrick Horain, IHCI 2014 Chair

11:00 – 12:30: S1.1: Vision-based interfaces

Mohamed Dahmane and Langis Gagnon
Local Phase-Context for Face Recognition under Varying Conditions

Sudhakar Mishra and Uma Shankar Tiwary
Heart Rate Measurement Using Video in Different User States for Online HCI Applications

Maxime Boucher, Fakhreddine Ababsa and Malik Mallem
On depth usage for a lightened visual SLAM in small environments

12:30 – 14:00: Lunch

14:00 – 15:00: Invited talk

Catherine Pelachaud
Interacting with socio-emotional agents

Score board

3RD ODI, ENGLAND IN SRI LANKA - ODI, 3 DECEMBER 2014 AT HAMBANTOTA

SRI LANKA — 242/8 (35.0 OV)

236/5 (33.4 OV) — **ENGLAND**

ENGLAND BEAT SRI LANKA BY 5 WICKETS (D/L METHOD)

MAN OF THE MATCH: JOS BUTTLER

Sri Lanka 242/8 in 35.0 Overs

England 236/5 in 33.4 Overs

	R	B	4s	6s	SR
Alastair Cook (C) c Kumar Sangakkara b Dhammika Prasad	34	42	5	0	80.95
Moeen Ali run out (Rangana Herath)	58	40	2	5	145.00
Alex Hales c Ajantha Mendis b Angelo Mathews	27	29	2	1	93.10
Joe Root not out	48	48	2	2	100.00
Ravi Bopara c Kumar Sangakkara b Rangana Herath	6	3	0	1	200.00

- Summary output may be a picture, a movie, an audio segment
- Likewise the input may be in these different multimedia forms
- Source information may be found from various sources

Types of summaries

- Objective

  Indicative vs. informative

- Relationship with the source document

  Extracts (representative paragraphs/sentences/phrases): “a summary consisting entirely of material copied from the input”

  Abstracts: “a concise summary of the central subject matter of a document” [Paice90].

- Context

  User-focused/Query-focused vs. Generic Summaries

- Generic summaries are aimed at a particular – usually broad – readership community
- Dimensions

  Single-document vs. multi-document

Ideal Summary

- One which allowed the subject to correctly guess all the salient ideas in the full-text of the source document

  - informative
  - Coherence
  - Salience

Parameters of Summarization System

- Compression Rate: Summary length/Source length
- Audience: User-focused vs. Generic
- Relation to Source: Extract vs. Abstract
- Function: Indicative vs. Informative
- Coherence: Coherent vs. Incoherent
- Span: Single vs. Multi-document

- Language: Monolingual or Multi-lingual or Cross-Lingual
- Genre
- Media

Human Summarization Process

- General process that humans use when summarizing written or spoken text can be describes as a three step process (Brandow 1995):

1. Understanding the content of the document
2. Identifying most important pieces of information
3. Rewriting this information

- We use operation such as deletion, generalization and compaction in this process
- We identify important information, delete nonessential information and then rewrite the remaining information to make it more general and more compact.

Example

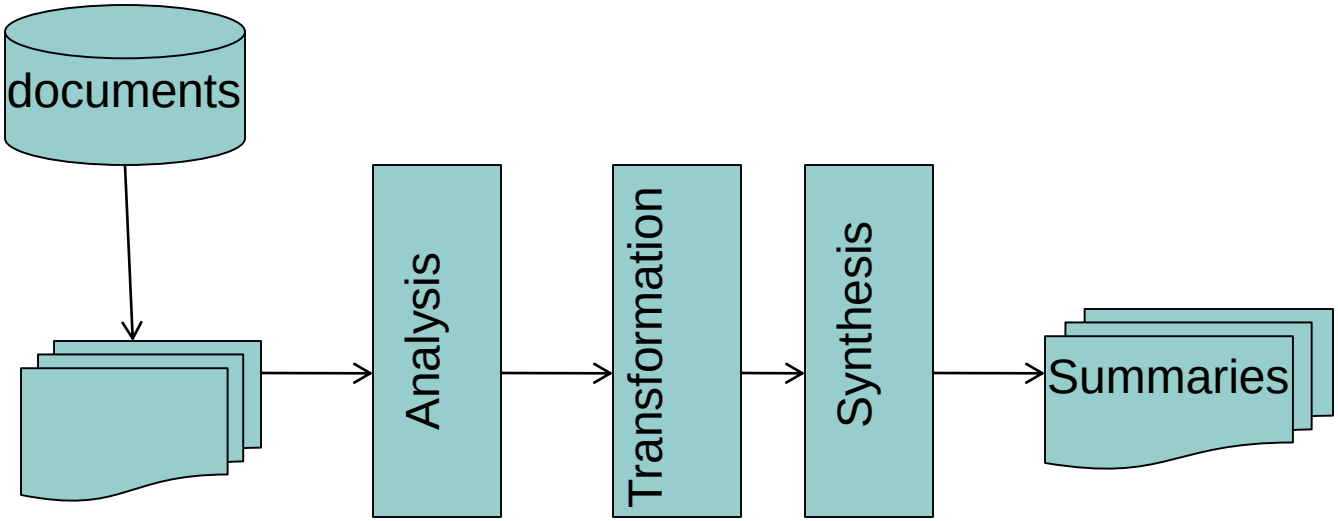
Yesterday morning my friend called me to visit her house. When I reached there my friend she was preparing coffee. Her father was cleaning dishes. Her mother was busy writing her new book.

- We can summarize the description of sample text by saying: *Yesterday when I visited my friend the whole family was busy*.

- Engres-Niggemeyer (1998) described the human summarization process using the following three stages:

1. Document exploration:
2. Relevance assessment:
3. Summary Production

Architecture for summarization



Single-document extracts: Analysis → Output

Professional Abstractors (Pinto Molina, 1995)

1. Interpretation
2. Selection
3. Reinterpretation
4. Synthesis

Methods/Approaches

- Shallow Approaches
- Deeper Approaches

Some of the Early Work

- Luhn (1958)
- Edmundson (1969)

Some Existing Summarizer Systems

- Autosummarize option in MS Office
- InXight Summarizer in the AltaVista
- IBM's Intelligent Miner
- DimSum Summarizer from SRA Corporation

Luhn(1958)

- Perhaps the most cited paper on summarization
- Proposed that frequency of a word in an articles provide a useful measure of its significance
- Significance factor was derived at sentence level and top ranking sentences were selected to form the auto abstract

Extraction : Edmondsonian Paradigm (1969)

- Features:
  - Cue words
  - Title Words
  - Keywords
  - Sentence Location

Sentence Weighting:

$$W(s) = \alpha\, C(s) + \beta\, K(s) + \gamma\, L(s) + \delta\, T(s)$$

Edmondson's Observations

- Best feature: location
- Worse Feature: Keywords
- Combination: Cue-title
- Evaluation was done on 200 scientific papers on Chemistry

Sentence Extraction as a Bayesian Classification (Kupiec et al, 1995)

Features used: sentence length, presence of fixed cue phrases, whether the sentence location was paragraph initial, paragraph-medial or paragraph-final, presence of thematic terms and presence of proper names

$$P(s \in E / F_1 F_2 ... F_n) = \frac{\prod_{i=1}^{n} P(F_i / s \in E) P(s \in E)}{\prod_{i=1}^{n} P(F_i)}$$

$P(s \in E)$ - Probability that a source sentence s is included in extract E

$P(F_i / s \in E)$ - Probability of feature F_i occurring in an extract sentence

- 188 full text/Summary pairs (Scientific Articles)
- Abstracts: written by professional abstractor (Average length 3 sentences)
- Best Individual Feature: Location
- Feature mix: location, cue phrase & Sentence length

Lin and Hovy (1997)

- Studied the importance of single feature, sentence position
- Underlying assumption: texts generally follow a predictable discourse structure & topic bearing sentences tend to occur in certain specifiable locations
- Corpus used: Newswire corpus

  - text about computer & related hardware + abstract of six sentences + a set of key topic words

- For each document in the corpus, yield of each sentence position against the topic keywords was computed
- Sentence positions were then ranked by their average yield to produce the Optimal Position Policy (OPP) for topic positions for the genre

Barzilay & Elhadad(1997)

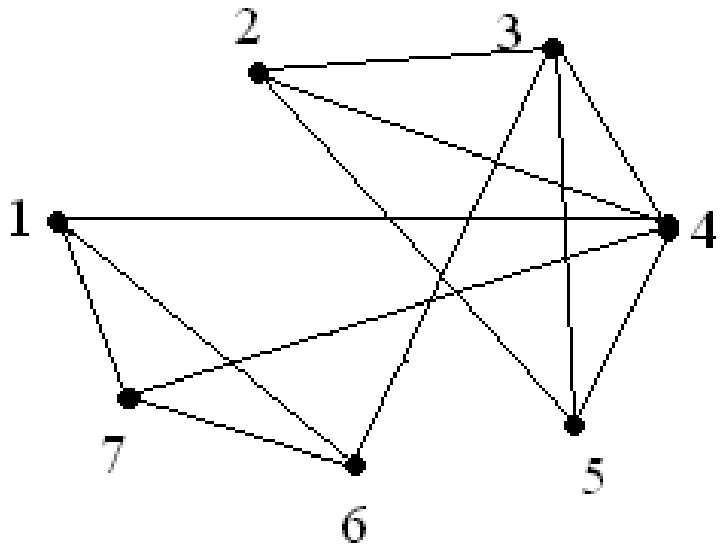
- Deep linguistic analysis

Steps:

1. Segmentation of text
2. Identification of lexical chains (sequence of related words): relatedness was measured in terms of WordNet distance
3. Using strong lexical chains to identify the sentences worthy of extraction

Graph-based Extraction (Salton, 1980)

- Graph-based methods map text into graphs.
- Nodes of the graph are textual units
- Two nodes are connected if they have vocabulary overlap above a threshold.
- Bushy nodes are good candidates for extraction


2
3
1
4
7
5
6

Summarization Evaluation

- *Intrinsic* approaches: assess the quality of a summary based on the analysis of the content of the summary itself
  1. Quality Evaluation
  2. Informativeness Evaluation
- *Extrinsic Approaches:* measure the summary based on how it affects the completion of certain tasks, e.g. IR

Intrinsic approaches

- Quality Evaluation:

  - to ask human judges to grade summaries for its *readability* or *acceptability*.

  - to automatically assess the quality of summaries using a grammar or style checker

- Informativeness is measured in terms of the amount of information preserved from the source text at different levels of compression or amount of information preserved from gold or ideal summary at different levels of compression

Sentence Recall and Sentence precision

Let m be the number of sentences in an ideal summary, n the number of sentences in a machine generated summary k of which also appear in the ideal summary.

$$SP = \frac{k}{n} \qquad SR = \frac{k}{m}$$

Utility-based measure

- (Radev et al 2000) uses a fine grained approach to judge summary worthiness of sentences.
- judges are asked to assign a score in between 1 to 10 to each sentence. These score are called utility points.
- the utility point of all the sentences in automatically generated summary that happen to be common with ideal summary are added up to evaluate the summary.

Content-based measures

- Content-based measures attempt to measure content similarity between a summary and its source
- can be used to evaluate both extracts as well as abstracts summary and the ‘gold’ summary.

Extrinsic Summary Evaluation

- Extrinsic summary evaluations assess the quality of a summary in terms of how it affects the performance of the task for which it has been generated

ROGUE (Recall Oriented Understudy for Gisting Evaluation)

- a software package for automated evaluation of summaries developed by Chin-Yew Lin ((USC)
- ROUGE summarization evaluation package (http://www.isi.edu/~cyl/ROUGE)
- Include the following automatic evaluation methods:

  -ROUGE-N: N-gram based co-occurrence statistics

  -ROUGE-L: LCS-based statistics

  -ROUGE-W: Weighted LCS that favors consecutive LCEes

  -ROUGE-S: Skip-bigram-based co-occurrence statistics

  -ROUGE-SU: Skip-bigram plus unigram-based co-occurrence statistic

ROUGE-W

Source string: [Overlap between consecutive pair of words]

Y: [Overlap between consecutive pair in text]

Z: [Overlap count between the consecutive pair in a document]

- ROUGE-L(Y) = ROUGE-L(Z)
- ROUGE-W(Y) > ROUGE-W(Z)

ROGUE-S: Skip-Bigram

- Any pair of words in their sentence order, allowing for arbitrary gaps.

Intuition:

- Consider long distance dependency.
- Allow gaps in matches as LCS but count all in-sequence pairs; while LCS only counts the longest subsequences.

ROGUE-S: Example

1. Overlap between consecutive pair
2. Overlap in consecutive pair
3. consecutive pair in overlap
4. consecutive pair overlap between

ROUGE-S: – S2=3/6 (“overlap consecutive”, “overlap pair”, “consecutive pair”)

– S3=1/6 (“consecutive pair”)

– S4=2/6 (“consecutive pair”, “Overlap between”)

– S2>S4>S3

ROGUE

R = { r1, r2, …, rn} be a set of reference summaries

S – automatic summary

$\varphi_n(d)$ – binary vector representing n-grams contained in d

$\varphi_n^i(d)$ = 1 if i-th n-gram is contained in d and 0 otherwise

$$ROGUE - N(s) = \frac{\sum_{r \in R} \langle \varphi_n(r), \varphi_n(s) \rangle}{\sum_{r \in R} \langle \varphi_n(r), \varphi_n(r) \rangle}$$

- Automatic and thus easy to apply
- Provides an opportunity to experiment with different units of comparison: uni-gram, bi-gram, LCS, skip-bigrams, basic elements

Multi document summarization (MDS)

- MDS is the process of filtering important information from a set of documents to produce a condensed version for particular users and application.
- It can be viewed as an extension of single document summarization.
- Issues like redundancy, novelty, coverage, temporal relatedness, compression ratio, etc., are more prominent in MDS (Radev et al 2004).

MDS: Issues

- Temporal ordering: In SDS sentence ordering is used
- Temporal references: References like “today”, “Tuesday”, can be ambiguous when extracted from multiple document
- Paraphrasing
- Repetition

Multi document summarization

- Input: set of documents
- Output: a summary containing important information across documents

- Pioneered by NLP group at Columbia University (McKeown and Radev, 1995) where SUMMONS (SUMMarizing Online NewS articles) was developed
- SUMMONS is an abstractive system that works in strict domain
- Relies on Template-driven IE Technology and NLG tools
- Targets single event in narrow domain

- DUC 2001, 2002, and 2003 MDS Tasks

  - Generic 10, 50, 100, 200 (2002) , and 400 (2001) words summaries

  - Short summaries of about 100 words in three different tasks in 2003

    » focused by an event (30 TDT clusters)

    » focused by a viewpoint (30 TREC clusters)

    » in response to a question (30 TREC Novelty track clusters)

MEAD (Radev *et al.* ,2004)

- MEAD is a large scale extractive system that works in general domain (http://www.summarization.com/mead)
- achieved good performance in large scale summarization of news articles

Input: cluster of documents with n sentences and compression rate (r)

Output: (n*r) sentences from the cluster with highest score

MEAD: feature set

Centroid: cosine overlap with the centroid vector of the cluster (Radev et al., 2004),

SimWithFirst: cosine overlap with the first sentence in the document (or with the title, if it exists),

Length: 1 if the length of the sentence is above a given threshold and 0 otherwise,

RealLength: the length of the sentence in words,

Position: the position of the sentence in the document,

QueryOverlap: cosine overlap with a query sentence or phrase,

KeyWordMatch: full match from a list of keywords,

LexPageRank: eigenvector centrality of the sentence on the lexical connectivity matrix with a defined threshold

- All features are computed on a sentence-by-sentence basis

MEAD Classifiers

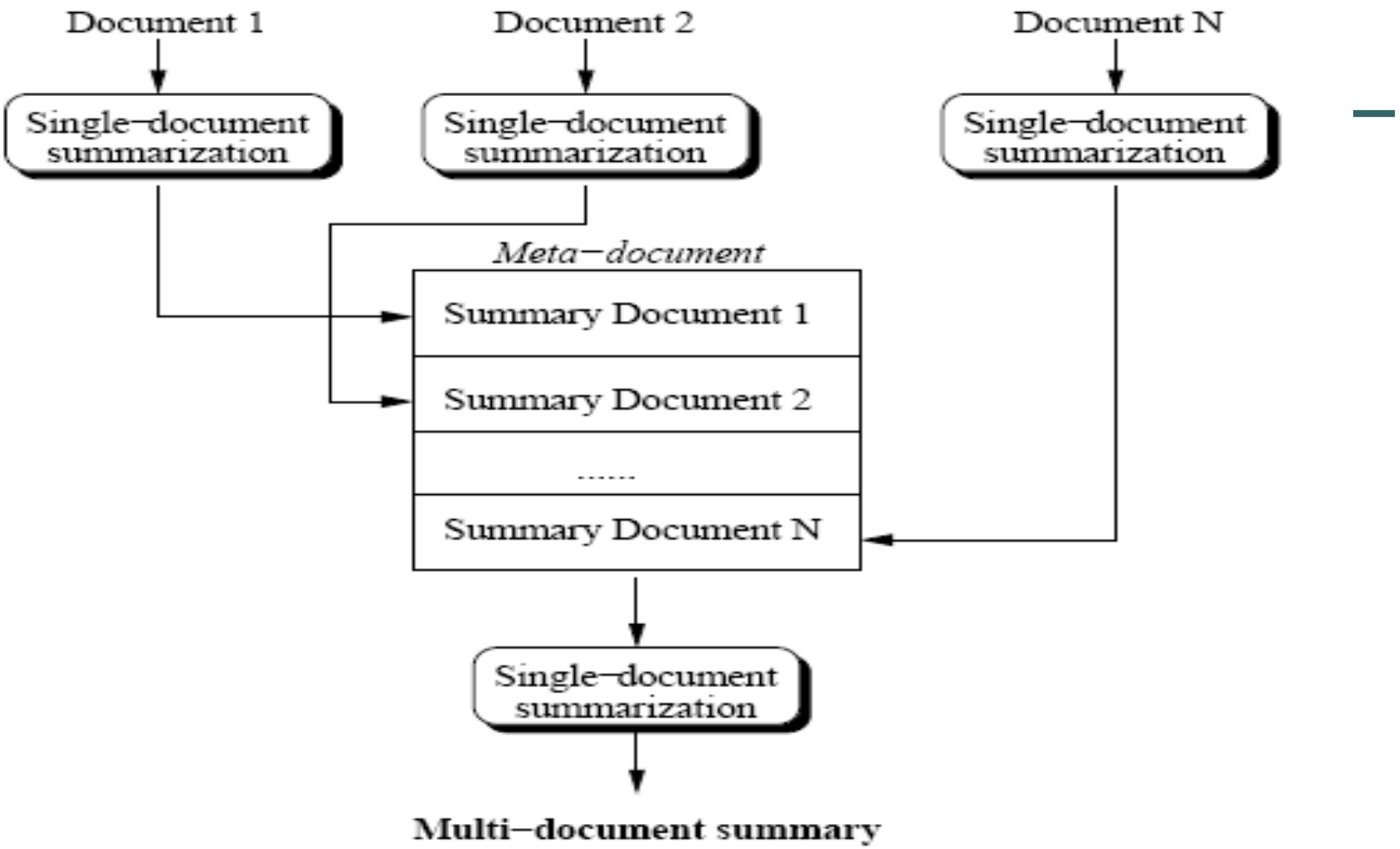
- Default: provides a linear combination of all features except for “Length” which is treated as a cutoff feature
- Lead-based: a baseline classifier that favors sentences that appear earlier in the cluster, as defined by the order of documents in the definition of the cluster,
- Random: a baseline classifier that extracts sentences at random from the cluster,
- Decision-tree: a machine learning algorithm, based on Weka (Witten and Frank, 2000) and trained on an annotated summary corpus.

MEAD evaluation toolkit

- toolkit allows evaluation of human-human, human-computer, and computer-computer agreement.
- Supports two general classes of evaluation metrics: co-selection and content-based metrics.
- Co-selection metrics include precision,recall, Kappa, and Relative Utility, a more flexible cousin of Kappa.
- MEAD's content-based metrics include cosine (which uses TF*IDF), simple cosine (which doesn't), and unigram- and bigram-overlap.

Graph-based algorithm for Extractive MDS (Mihalcea and Tarau. 2005)

- Language Independent Extractive Summarization based on graph-ranking algorithms
- Layered application of single document summarization
- PageRank and HITS ranking algorithms are used



MDS using Meta-Summarization(Mihalcea and Tarau(2005)

Single Document Summary Generation

1. Construct a graph by adding one vertex for each sentence
2. Establish edges using sentence-interconnections (defined using similarity)
3. Graph can be undirected, directed forward or directed backward
4. Run ranking algorithm and sort sentences in reverse order of their score
5. Select top ranked sentences for inclusion in extractive Summary

Let $G = (V, E)$ be a directed graph with the set of vertices V and set of edges E, where E is a subset of $V \times V$. For a given vertex V_i, let $In(V_i)$ be the set of vertices that point to it (predecessors), and let $Out(V_i)$ be the set of vertices that vertex V_i points to (successors).

$$PR(V_i) = (1 - d) + d * \sum_{V_j \in In(V_i)} \frac{PR(V_j)}{|Out(V_j)|}$$

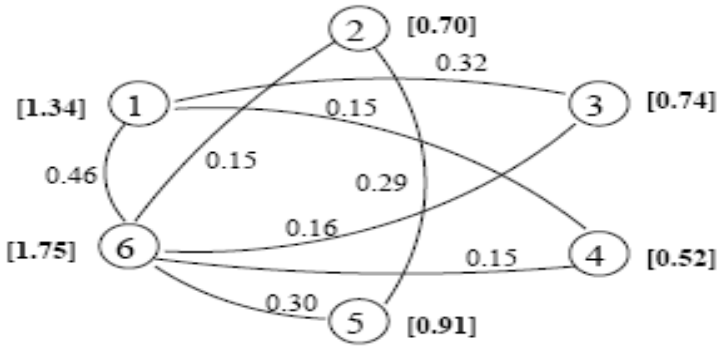
where d is a parameter set between 0 and 1.

- PageRank algorithm is to be adapted for weighted graph.

$$PR^W(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} w_{ji} \frac{PR^W(V_j)}{\sum_{V_k \in Out(V_j)} w_{kj}}$$

Example (Mihalcea & Tarau (2005)

[1] Watching the new movie, "Imagine: John Lennon," was very painful for the late Beatle's wife, Yoko Ono.
[2] "The only reason why I did watch it to the end is because I'm responsible for it, even though somebody else made it," she said.
[3] Cassettes, film footage and other elements of the acclaimed movie were collected by Ono.
[4] She also took cassettes of interviews by Lennon, which were edited in such a way that he narrates the picture.
[5] Andrew Solt ("This Is Elvis") directed, Solt and David L. Wolper produced and Solt and Sam Egan wrote it.
[6] "I think this is really the definitive documentary of John Lennon's life," Ono said in an interview.



Multi-document Summary Generation

- Generate summary for each document in a given cluster of documents using one of the graph-based ranking algorithms
- Produce a “summary of summaries” is using the same or a different ranking algorithm
- a maximum threshold on sentence similarity measure is used to avoid pair of sentences having highly similar content while constructive graph for creating “summary of summaries”.

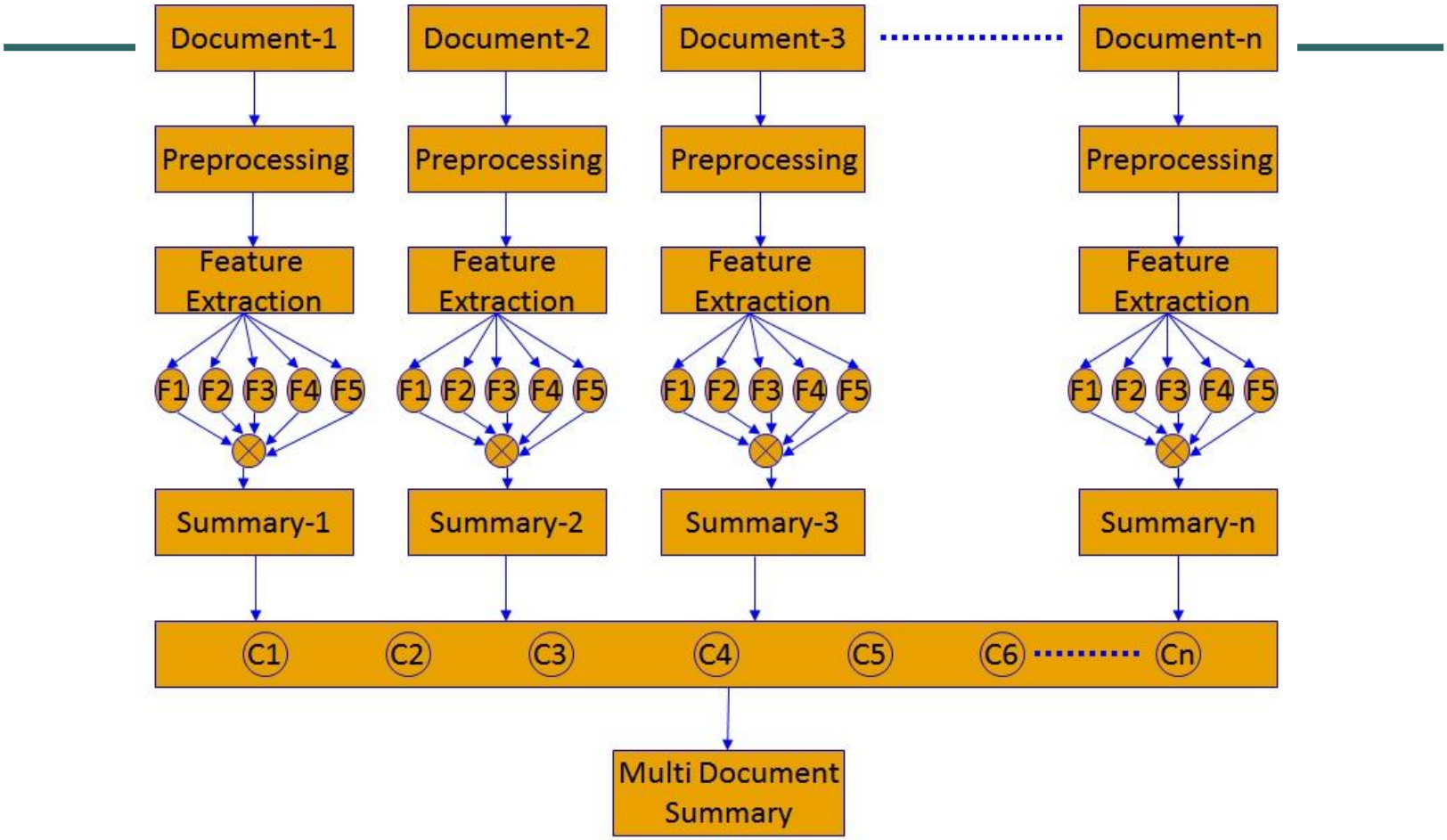
Evaluation

- Single Document Summary: 100 word summaries for each of the 567 English news articles provided during the Document Understanding Evaluations 2002 and TeM´ario data set (Portuguese)
- Multi-Document Summary: DUC 2002 dataset
- The results are comparable to top 5 best performing summarization systems in DUC 2002

Multi Document Summarization Using Sentence Clustering (Gupta & Siddiqui, 2012)

- Combines single document summaries using sentence clustering
- Uses syntactic and semantic similarity between sentence for clustering
- DUC 2002 multi-document dataset for evaluation

MDS using Sentence Clustering



Algorithm

Steps :

1. Preprocessing
2. Feature Extraction
3. Single Document Summary Generation
4. Multi Document Summary Generation

Preprocessing

- Noise Removal
- Tokenization
- Stop word Removal
- Stemming
- Frequency Analysis
- Sentence splitting

Feature Extraction

- Document Feature
- Location Feature
- Sentence Reference Index Feature
- Concept Similarity Feature

Single Document Summary Generation

1. Calculate Sentence weight:

$$S(W)=u*D(f)+v*L(f)+w*SRI(f)+x*CS(f).$$

2. Normalize sentence weight

3 Extract top k sentences

Multi-Document Summarization

- Take individual document summaries and create sentence clusters
- Extract sentences from each cluster.
- Arrange the extracted sentences on the basis of position in the original document.

Syntactic Similarity (Li et al., 2006)

$$Sim_0(S_1, S_2) = \frac{\sum(v_0 * v_r) - \frac{\sum v_0 * \sum v_r}{k}}{\sqrt{(\sum {v_0}^2 - \frac{(\sum v_0)^2}{k})(\sum {v_r}^2 - \frac{(\sum v_r)^2}{k})}}$$

Where, k is the no. of words in sentence S1.

Vo is Original Order Vector

Vr is Relative Order Vector

Semantic Similarity(Li et al., 2006)

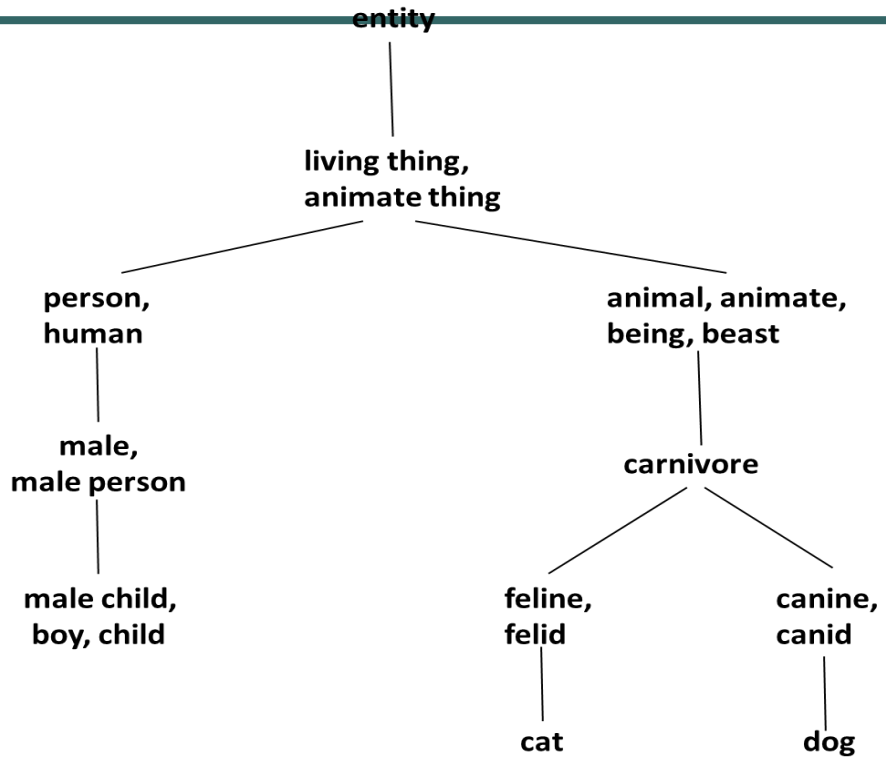


Fig. 1. A part of WordNet-style hierarchy

- Shortest Path Length(l)
- Depth of Subsumer(d)

$$S_w(w_1, w_2) = \frac{f(d)}{f(d) + f(l)}$$

$$f(x) = e^{\alpha x} - 1$$

where, α is a smoothing factor

- Calculate information content of each word in a corpus (BNC)

Semantic Similarity between S1 and S2 is (Li et al., 2006):

$$Sim_s(S_1, S_2) = \frac{\sum_{w_i \in S_1} \max_{w_j \in S_2} \left(S_w(w_i, w_j) * I_{w_i}\right)}{\sum_{w_i \in S_1} I_{w_i} + \sum_{w_j \in S_2} I_{w_j}}$$

$S_w(w_1, w_2)$is the semantic similarity between words.

Overall Sentence Similarity

- The overall similarity between two sentences, S1 and S2 is calculated as (Liu et al. 2008):

$$Sim_{sen} = Sim_s(S_1, S_2) * \big((1-\gamma) + \gamma * Sim_0(S_1, S_2)\big) + Sim_s(S_2, S_1) * \big((1-\gamma) + \gamma * Sim_0(S_2, S_1)\big)$$

Where γ is a smoothing factor.

Multi Document Summary

1. Extract sentences from each cluster.
2. Arrange the extracted sentences according to their position in the original document.

Evaluation

Dataset: DUC 2002, 100 word gold standard summary
Performance Measures: Recall, Precision and F-measure

Table 1: Results of Single Document Summarization

Average Recall	**0.45947**
Average Precision	**0.47989**
Average F-Measure	**0.46768**

Table 2: Results of MDS using Sentence Clustering

Average Recall	**0.33358**
Average Precision	**0.34221**
Average F-Measure	**0.33774**

Table 3: DUC 2002 Best Results

Top 5 Systems (DUC 2002)					
S26	S19	S29	S25	S20	Baseline
0.3578	0.3447	0.3264	0.3056	0.3047	0.2932

Language generation approach (Barzilay et al.,)

- Automatically generate a summary by identifying and synthesizing similar elements from a set of related documents
- For Generation FUF/SURGE system was used.

Algorithmic Steps (Barzilay,….)

1. Content selection to identify theme Intersection: a predicate argument structure was used to identify common parts
2. Corpus Analysis for deriving Paraphrasing rules: ordering of sentence component, main clause vs. relative clause, Realization in different syntactic categories, (Pentagon Speaker vs speaker from pentagon) change in grammatical feature, head omission

Algorithmic Steps (Barzilay,….)

3. Temporal ordering: approximation by with publication, substitute temporal ref date with full time/date reference
4. Sentence Generation: Input phrases are mapped to a SURGE syntactic input and sentences are generated

References

H P. Luhn, The automatic creation of literature abstracts, IBM Journal of Research and development archive Volume 2 Issue 2, April 1958, Pages 159-165.

H. P. Edmundson, New methods in automatic extracting, Journal of the ACM (JACM) JACM Homepage archive Volume 16 Issue 2, April 1969, Pages 264-285.

Dipanjan Das , André F. T. Martins , A Survey on Automatic Text Summarization (2007)

K. McKeown and D. Radev, "Generating summaries of multiple news articles", In Proceedings of the 18th Annual International ACM , (pp.74-82). Seattle, WA, 1995.

Regina Barzilay and Michael Elhadad, "Using Lexical chains for Text Summarization", ACL/EACL Workshop on Intelligent Scalable Text Summarization, pages 10-17, Madrid, 1997.

Conference on Research and Development in Information Retrieval (ACM SIGIR) (pp.74-82). Seattle, WA, 1995.

D. Radev, Hongyan Jing, Malgorzata Stys and Daniel Tam, "Centroid-based summarization of multiple documents", Information Processing and Management 40 919–938, 2004.

Dragomir R. Radev, Hongyan Jing, Malgorzata Stys and Daniel Tam, "Centroid-based summarization of multiple documents", Information Processing and Management 40 919–938, 2004.

Yuhua Li, David McLean, Zuhair A. Bandar, James D. O'Shea, and Keeley Crockett, "Sentence Similarity Based on Semantic Nets and Corpus Statistics", IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No. 8, August 2006, p. 1138-1150.

Yuhua Li, David McLean, Zuhair A. Bandar, James D. O'Shea, and Keeley Crockett, "Sentence Similarity Based on Semantic Nets and Corpus Statistics", IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No. 8, August 2006, p. 1138-1150.

Alkesh Patel, Tanveer J Siddiqui and U S Tiwary, "A language independent approach to Multi-lingual Text Summarization", In the proceedings of RIAO 2007, May 30 to June 1, 2007. Available at: http://riao.free.fr/papers/30.pdf

British National Corpus [Online]. Available: http://www.natcorp.ox.ac.uk/

Virendra Gupta & Tanveer J. Siddiqui, Multi-Document Summarization Using Sentence Clustering, IEEE Proceedings of 4th International Conference on Intelligent Human Computer Interaction, Kharagpur, India, December 27-29, 2012, p314-318.

Queries ?

Thanks