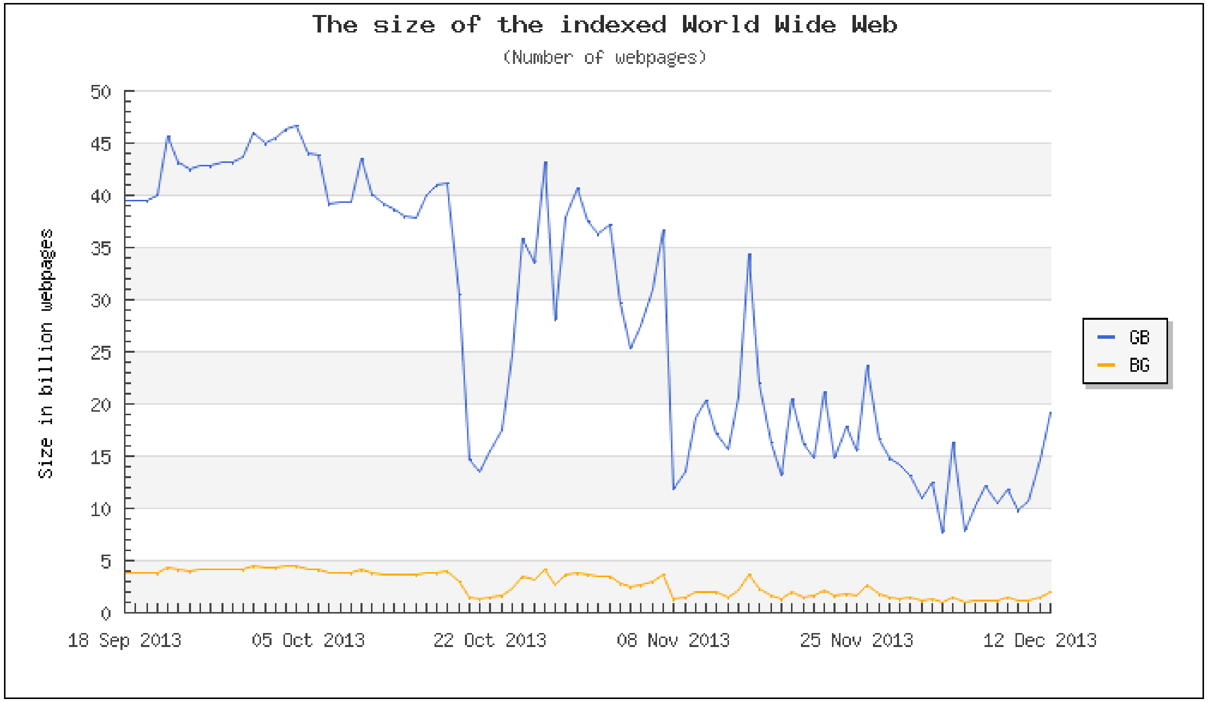
Multi-Document Summarization (MLTA 2013)

Dr. Tanveer J. Siddiqui

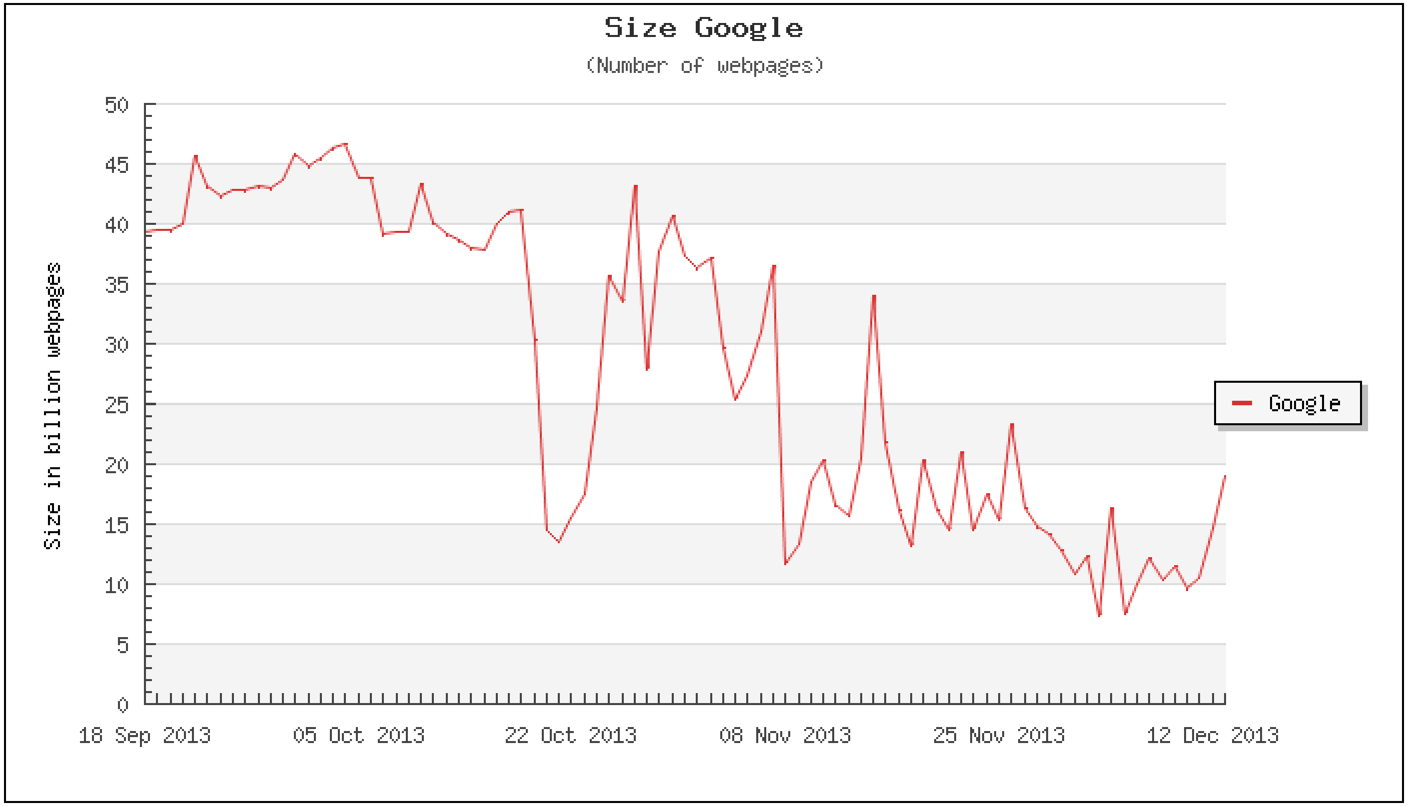
Outline

- 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

Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization

Rada Mihalcea
Department of Computer Science
University of North Texas
rada@cs.unt.edu

Abstract

This paper presents an innovative unsupervised method for automatic sentence extraction using graph-based ranking algorithms. We evaluate the method in the context of a text summarization task, and show that the results obtained compare favorably with previously published results on established benchmarks.

1 Introduction

Graph-based ranking algorithms, such as Kleinberg's HITS algorithm (Kleinberg, 1999) or Google's PageRank (Brin and Page, 1998), have been traditionally and successfully used in citation analysis, social networks, and the analysis of the link-structure of the World Wide Web. In short, a graph-based ranking al-

algorithms – previously found to be successful on a range of ranking problems. We also show how these algorithms can be adapted to undirected or weighted graphs, which are particularly useful in the context of text-based ranking applications.

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).

2.1 HITS

HITS (Hyperlinked Induced Topic Search) (Kleinberg, 1999) is an iterative algorithm that was designed

Movie trailer

User Rating

138 Votes (3.58)

The Amazing Spider-Man 2
(Columbia Pictures)

Release Date: May 2, 2014

Director: Marc Webb
Writer: Alex Kurtzman, Roberto Orci, Jeff Pinkner, James Vanderbilt
Cast: Andrew Garfield, Emma Stone, Jamie Foxx, Shailene Woodley, Dane DeHaan, Colm Feore, Paul Giamatti, Sally Field, Chris Cooper, B.J. Novak, Sarah Gadon

Plot: A sequel to the 2012 blockbuster that follows the continuing adventures of Peter Parker, also known as Spider-Man.

Genre: Action, Adventure, Fantasy

IMDb: tt1872181
Website: www.theamazingspiderman.com

Tentative Schedule

(15^{th} – 23^{rd} Dec. 2013)

Pre-Workshop Tutorials

Day/ Time	9:30 AM – 11:30 AM	12:00 Noon – 2:00 PM	3:00 PM – 5:00 PM
Sunday 15.12.2013	Danish Lohani	J R Bhatnagar	Manoj Singh

Keynotes/ Tutorials/ Short Talks

Day/ Time	9:30 AM – 11:00 AM	11:30 AM – 1:00 PM	2:00 PM – 3:30 PM	4:00 PM – 5: 30 PM
Monday 16.12.2013	Registration & Inaugural		David Barber	
Tuesday 17.12.2013	Pushpak Bhattacharya	Rakesh Agrawal	Radhika Mamidi	
Wednesday 18.12.2013	David Barber		Paper Presentations	
Thursday 19.12.2013	Deepayan Sarkar		Bing Liu	Madhu
Friday 20.12.2013	Alexandar Gelbukh	Vivek Singh	Jayadeva	
Saturday 21.12.2013	Indrajit Bhattacharya	T J Siddiqui	Niladri Chatterjee	
Sunday 22.12.2013	Asif Ekbal		P K Singh	Evaluation & Feedback
Monday 23.12.2013	Ganesh Ramakrishnan		Valedictory	

Score board

- 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].

Extract vs. Abstract

Many languages have changed and developed because of outside influences (1). English as we know it today, for example, has many words adapted from other cultures (2). It has some Latin words from the days when it was part of Roman Empire (3). English has a large number of words derived from French, the language of England's ruling classes following the Norman invasion of 1066 (4). Spanish Italian, French, Portuguese and Romanian languages all have many similar words (5). This is because they are descendent from Latin (6). Latin is the language of Roman Empire, of which Spain, Italy, France, Portugal and Romania were once part. (7)

Extract

Spanish Italian, French, Portuguese and Romanian languages all have many similar words (5). Latin is the language of Roman Empire, of which Spain, Italy, France, Portugal and Romania were once part. (7)

Abstract

Many languages have changed and devolved because of outside influence including English which has many words from Latin and Roman. Spanish, Italian, French, Portuguese and Romanian languages all have many similar words as they were once part of Roman Empire.

- 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

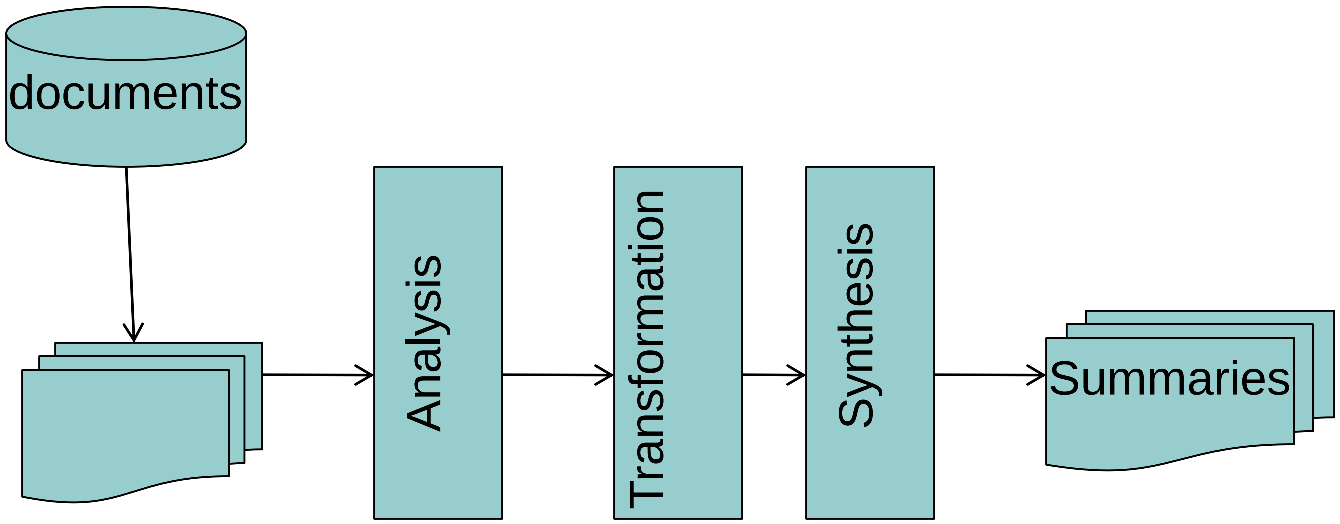
Yesterday morning my friend called me to visit her house. When I reached there my friend she was preparing coffee. Her father was cleaning dishes. Her mother was busy writing her new book.

- 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

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

$P(s \in E)$ - 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)

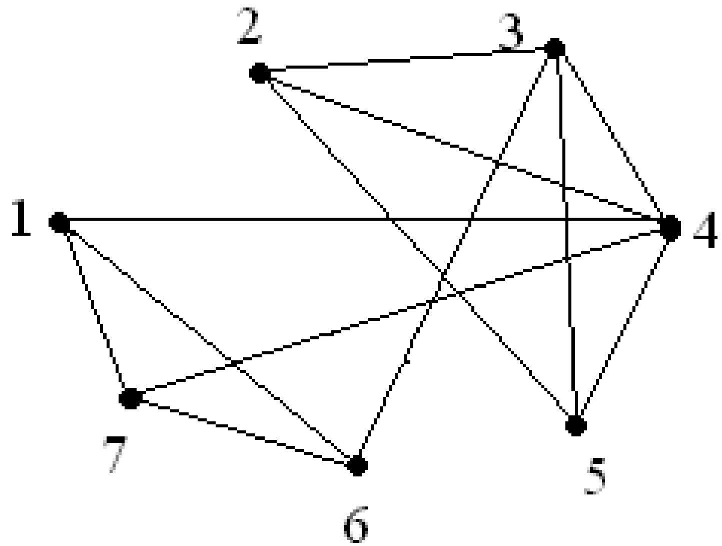
- Deep linguistic analysis

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
6
5

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

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

- Informativess 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)

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

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

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).

- 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

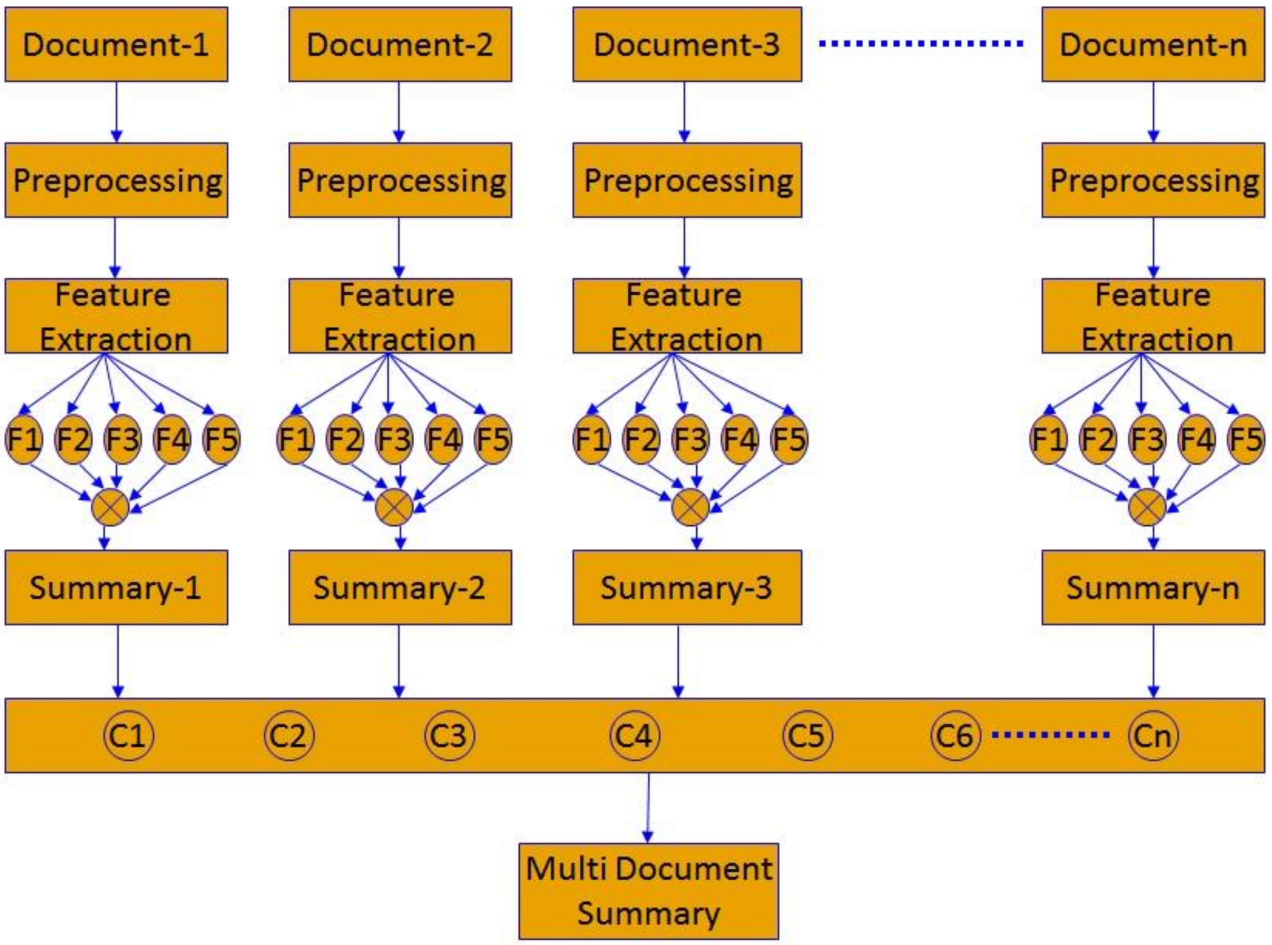
MEAD

- MEAD is a large scale extractive system that works in general domain
- achieved good performance in large scale summarization of news articles

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)

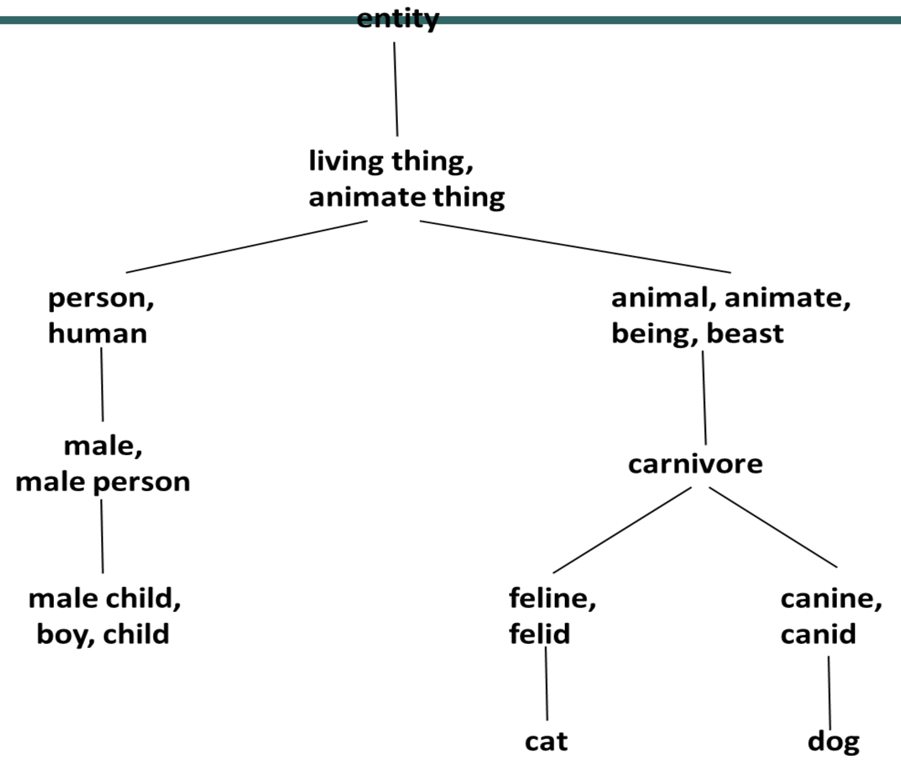


Fig. 1. A part of WordNet-style hierarchy

- 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

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: www.natcorp.ox.ac.uk/

Virendra Gupta http://& 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.

Karen Sparck-Jones, Automatic summarizing: The state of the art, Information Processing and Management : Volume 43 Issue 6, November, 2007, Pages 1449-1481.

Gerard Salton , Amit Singhal , Mandar Mitra , Chris Buckley, Automatic text structuring and summarization, Information Processing and Management: an International Journal, v.33 n.2, p.193-207, March 1997

Sparck-Jones, Automatic summarising: Factors and directions. In: Mani, I., Maybury, M.T. (Eds.), A dvances in automatic text summarisation, MIT Press, Cambridge, MA. pp. 1-14. 1997

http://www.summarization.com/mead/