Tutorial Programme

Tutorial Organiser

Horacio Saggion
Department of Computer Science
University of Sheffield
211 Portobello Street
Sheffield – S1 4DP
United Kingdom

Introduction to Text Summarization and Other Information Access Technologies

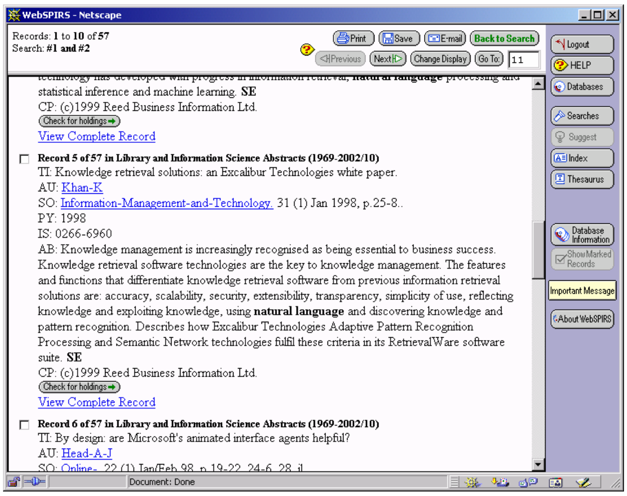
Horacio Saggion
Department of Computer Science
University of Sheffield
England, United Kingdom
saggion@dcs.shef.ac.uk

Automatic Text Summarization

- An information access technology that given a document or sets of *related* documents, extracts the most important content from the source(s) taking into account the user or task at hand, and presents this content in a well formed and concise text

LREC 2008 – Marrakech - Morocco

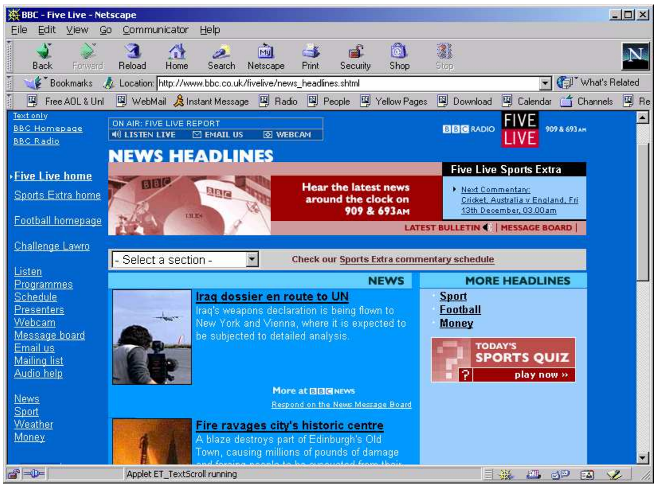
Examples of summaries – abstract of research article



3

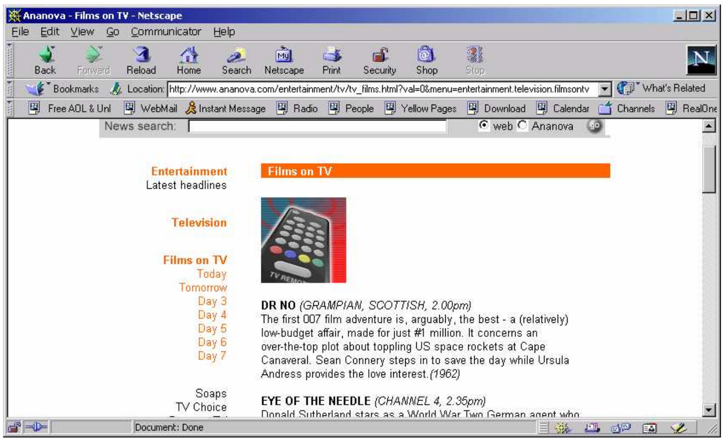
LREC 2008 – Marrakech - Morocco

Examples of summaries – headline + leading paragraph



4

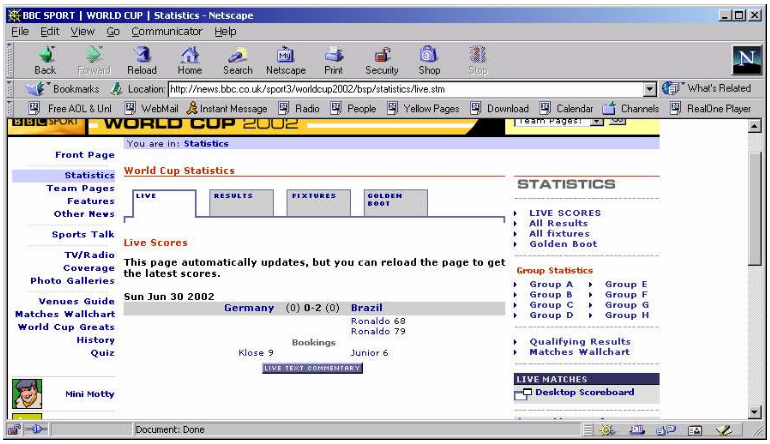
Examples of summaries – movie preview



5

LREC 2008 – Marrakech - Morocco

Examples of summaries – sports results



6

Summarization Parameters

- input document or document cluster
- compression: the amount of text to present or the length of the summary to the length of the source.
- type of summary: indicative/informative/... abstract/extract…
- other parameters: topic/question/user profile/...

What is a summary for?

- Direct functions
  - communicates substantial information;
  - keeps readers informed;
  - overcomes the language barrier;
- Indirect functions
  - classification; indexing; keyword extraction; etc.

Typology

- Indicative
  - indicates types of information
  - "alerts"
- Informative
  - includes quantitative/qualitative information
  - "informs"
- Critic/evaluative
  - evaluates the content of the document

ATTENTION: Earthquake in Turkey!!!!

Earthquake in the town of Cat in Turkey. It measured 5.1 in the Richter scale. 4 people dead confirmed.

Earthquake in the town of Cat in Turkey was the most devastating in the region.

Indicative/Informative distinction

INDICATIVE

The work of Consumer Advice Centres is examined. The information sources used to support this work are reviewed. The recent closure of many CACs has seriously affected the availability of consumer information and advice. The contribution that public libraries can make in enhancing the availability of consumer information and advice both to the public and other agencies involved in consumer information and advice, is discussed.

INFORMATIVE

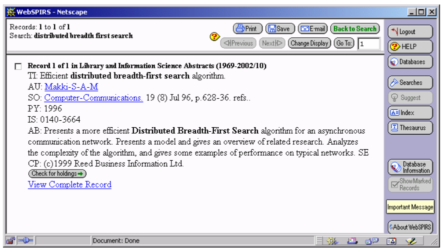
An examination of the work of Consumer Advice Centres and of the information sources and support activities that public libraries can offer. CACs have dealt with pre-shopping advice, education on consumers' rights and complaints about goods and services, advising the client and often obtaining expert assessment. They have drawn on a wide range of information sources including case records, trade literature, contact files and external links. The recent closure of many CACs has seriously affected the availability of consumer information and advice. Libraries can cooperate closely with advice agencies through local coordinating committed, shared premises, join publicity referral and the sharing of professional expertise.

More on typology

- extract vs abstract
  - fragments from the document
  - newly re-written text
- generic vs query-based vs user-focused
  - all major topics equal coverage
  - based on a question "what are the causes of the war?"
  - users interested in chemistry
- for novice vs for expert
  - background
  - Just the new information
- single-document vs multi-document
  - research paper
  - proceedings of a conference
- in textual form vs items vs tabular vs structured
  - paragraph
  - list of main points
  - numeric information in a table
  - with "headlines"
- in the language of the document vs in other language
  - monolingual
  - cross-lingual

Abstracting services

- Abstracting journals
  - not very popular today
- Abstracting databases
  - CD-ROM
  - Internet
- Mission
  - keep the scientific community informed
- LISA, CSA, ERIC, INSPEC, etc.
- employ professional abstractors



Transformations during abstracting

Cremmins: The art of abstracting

Source document	Abstract
There were significant positive associations between the concentration of the substance administered and mortality in rats and mice of both sexes.	Mortality in rats and mice of both sexes was dose related.
There was no convincing evidence to indicate that endrin ingestion induced any of the different types of tumors which were found in the treated animals.	No treatment related tumors were found in any of the animals.

Study of professional abstractors/abstracts

- Abstractor's at work (Endres-Niggemeyer'95)
- Abstract's structure (Liddy'91)
- What information from documents is used to create abstracts (Saggion&Lapalme'02)

Automatic Summarization

- 50s-70s
  - Statistical techniques (scientific text)
- 80s
  - Artificial Intelligence (short texts, narrative, some news)
- 90s-
  - Hybrid systems (news, some scientific text)
- 00s-
  - Headline generation; multi-document summarization (much news, more diversity: law, medicine, e-mail, Web pages, etc.); hand-held devices; multimedia

Summarization and other information access technologies

- Information Retrieval
  - open domain, given a "query" returns documents matching the query
  - summaries can provide access points for quick check document relevance specially if they take into consideration the user query
- Information Extraction
  - domain dependent, given a "template" instantiate its slots with "strings" from the document
  - template represents the key information of an event
  - Domain specific summaries can be created from the template

Summarization and other information access technologies

- Question answering
  - open domain, given a well formed natural language question and a collection of documents, returns answers to the question
  - summarization can be used to present the answers
  - definitions/profiles usually required in QA settings are specific types of summaries

Summarization steps

- Text interpretation
  - phrases; sentences; propositions; etc.
- Unit selection
  - some sentences; phrases; props; etc.
- Condensation
  - delete duplication, generalization
- Generation
  - text-text; propositions to text; information to text

Natural language processing to support summarization

detecting syntactic structure for condensation

I: Solomon, a sophomore at Heritage School in Convers, is accused of opening fire on schoolmates.

O: Solomon is accused of opening fire on schoolmates.

meaning to support condensation

I: 25 people have been killed in an explosion in the Iraqi city of Basra.

O: Scores died in Iraq explosion

discourse interpretation/coreference

I: And as a conservative Wall Street veteran, Rubin brought market credibility to the Clinton administration.

O: Rubin brought market credibility to the Clinton administration.

I: Victoria de los Angeles died in a Madrid hospital today. She was the most acclaimed Spanish soprano of the century. She was 81.

O: Spanish soprano De los Angeles died at 81.

LREC 2008 – Marrakech - Morocco

Summarization by sentence extraction

- extract
  - subset of sentence from the document
- easy to implement and robust
- how to discover what type of linguistic/semantic information contributes with the notion of relevance?
- how extracts should be evaluated?
  - create ideal extracts
  - need humans to assess sentence relevance

Evaluation of extracts

choosing sentences

N	Human	System
1	+	+
2	-	+
n	-	-

contingency table

	S	
H	+	-
+	TP	FN
-	FP	TN

True Positive

False Positive

False Negative

True Negative

- precision $\frac{TP}{TP+FP}$
- recall $\frac{TP}{TP+FN}$

$$TP + FN + TN + FP = n$$

LREC 2008 – Marrakech - Morocco

Evaluation of extracts (instance)

N	Human	System
1	+	+
2	-	+
3	+	-
4	-	-
5	+	-

	S	
H	+	-
+	1	2
-	1	1

- precision = 1/2
- recall = 1/3

Summarization by sentence scoring and ranking

- Document = set of sentences S
- Features = set of features F
- For each sentence S_k in the document
  - For each feature F_i
    - V_i = compute_feature_value(S_k, F_i)
  - $score_k$= combine_features(F);
- Sorted = Sort (< S_k, $score_k$>) in descending order of $score_k$
- Select top ranked m sentences from Sorted
- Show sentences in document order

Superficial features for summarization

- Keyword distribution (Luhn'58)
- Position Method (Edmundson'69)
- Title Method (Edmundson'69)
- Cue Method/Indicative Phrases (Edmundson'69; Paice'81)

Some details

- Keyword = a word "statistically" significant according to its distribution in document/corpus
  - each word gets a score
  - sentence gets a score (or value) according to the scores of the words it contains
- Title = a word from title
  - sentence gets a score according to the presence of title words

Some details

- Cue = there is a predefined list of words with associated weights
  - associate to each word in a sentence its weight in the list
  - score sentence according to the presence of cue words
- Position = sentences at beginning of document are more important
  - associate a score to each sentence depending on its position in the document

Experimental combination (Edmundson'69)

- Contribution of 4 features
  - title, cue, keyword, position
  - linear equation

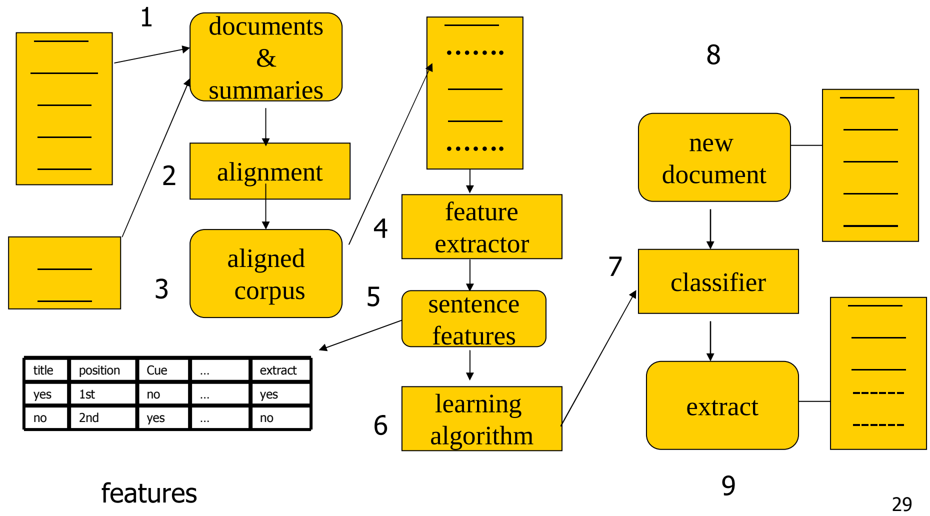
$$Weight(S) = \alpha.Title(S) + \beta.Cue(S) + \gamma.Keyword(S) + \delta.Position(S)$$

  - first the parameters are adjusted using training data

Experimental combination

- All possible combinations 4^2 - 1 (=15 possibilities)
  - title + cue; title; cue; title + cue + keyword; etc.
- Produces summaries for test documents
- Evaluates co-selection (precision/recall)
- Obtains the following results
  - best system
    - cue + title + position
  - individual features
    - position is best, then
    - cue
    - title
    - keyword

Learning to extract



LREC 2008 – Marrakech - Morocco

Statistical combination

- method adopted by Kupiec&al'95
- need corpus of documents and extracts
  - professional abstracts
- alignment
  - program that identifies similar sentences
  - manual validation

Statistical combination (features)

- length of sentence (true/false)

$$len(S) > u_l$$

- cue (true/false)

$$(S_i \cap DIC_{cue}) \neq \phi$$

or

$$heading(S_{i-1}) \wedge (S_{i-1} \cap DIC_{headings}) \neq \phi$$

LREC 2008 – Marrakech - Morocco

Statistical combination (features)

- position (discrete)
  - paragraph # $\{1,2,...,10\} \vee \{last, last-1,..., last-4\}$
  - in paragraph $\{initial, middle, final\}$
- keyword (true/false) $rank(S) > u_k$
- proper noun (true/false)
  - similar to keyword

Statistical combination

- combination

features in
extract sentences

sentence belongs
to extract given features

prob. of
sentence
in extract

Bayes theorem

$$p(s \in E | f_1, \ldots, f_n) = \frac{p(f_1, \ldots, f_n | s \in E).p(s \in E)}{p(f_1, \ldots, f_n)}$$

features in
corpus

Statistical combination

- parameter
  estimation

assume
independence

$$p(f_1, \ldots, f_n | s \in E) = \prod p(f_i | s \in E)$$

$$p(f_1, \ldots, f_n) = \prod p(f_i)$$

estimate
by counting

$$p(s \in E)$$

LREC 2008 – Marrakech - Morocco

Statistical combination

- results for individual features
  - position
  - cue
  - length
  - keyword
  - proper name
- best combination
  - position+cue+length

35

LREC 2008 – Marrakech - Morocco

Problems with extracts

- Lack of cohesion

source

A single-engine airplane crashed Tuesday into a ditch beside a dirt road on the outskirts of Albuquerque, killing all five people aboard, authorities said.

Four adults and one child died in the crash, which witnesses said occurred about 5 p.m., when it was raining, Albuquerque police Sgt. R.C. Porter said.

The airplane was attempting to land at nearby Coronado Airport, Porter said.

It aborted its first attempt and was coming in for a second try when it crashed, he said…

extract

Four adults and one child died in the crash, which witnesses said occurred about 5 p.m., when it was raining, Albuquerque police Sgt. R.C. Porter said.

It aborted its first attempt and was coming in for a second try when it crashed, he said.

36

Problems with extracts

- Lack of coherence

source

Supermarket A announced a big profit for the third quarter of the year. The directory studies the creation of new jobs. Meanwhile, B's supermarket sales drop by 10% last month. The company is studying closing down some of its stores.

extract

Supermarket A announced a big profit for the third quarter of the year. The company is studying closing down some of its stores.

LREC 2008 – Marrakech - Morocco

Approaches to cohesion

- identification of document structure
- rules for the identification of anaphora
  - pronouns, logical and rhetorical connectives, and definite noun phrases
  - Corpus-based heuristics
- aggregation techniques
  - IF sentence contains anaphor THEN include preceding sentences
- anaphora resolution is more appropriate but
  - programs for anaphora resolution are far from perfect

Approaches to cohesion

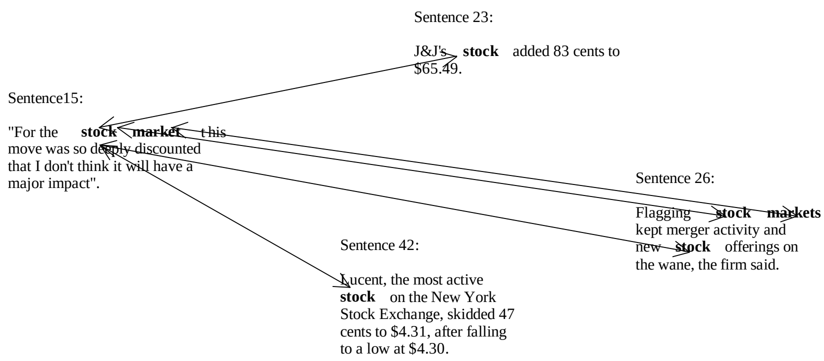
- BLAB project (Johnson & Paice'93 and previous works by same group)
  - rules for identification: "that" is :
    - non-anaphoric if preceded by research-verb (e.g. "assume", "show", etc.)
    - non-anaphoric if followed by pronoun, article, quantifier, demonstrative,…
    - external if no latter than 10^{th} word of sentence
    - else: internal
  - selection (indicator) & rejection & aggregation rules; reported success: abstract > aggregation > extract

Telepattan system: (Bembrahim & Ahmad'95)

- Link two sentences if
  - they contain words related by repetition, synonymy, class/superclass (hypernymy), paraphrase
    - *destruct* ~ *destruction*
  - use thesaurus (i.e., related words)
- pruning
  - links(s_i, s_j) > thr => bond (s_i, s_j)

LREC 2008 – Marrakech - Morocco

Telepattan system

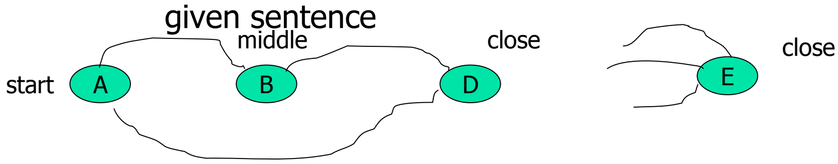


41

LREC 2008 – Marrakech - Morocco

Telepattan system

- Classify sentences as
  - start topic, middle topic, end of topic, according to the number of links
  - this is based on the number of links <u>to</u> and <u>from</u> a given sentence



- Summaries are obtained by extracting sentences that open-continue-end a topic

42

Lexical chains

- Lexical chain:
  - word sequence in a text where the words are related by one of the relations previously mentioned
- Use:
  - ambiguity resolution
  - identification of discourse structure
- Wordnet Lexical Database
  - synonymy: dog, can
  - hypernymy: dog, animal
  - antonym: dog, cat
  - meronymy (part/whole): dog, leg

43

LREC 2008 – Marrakech - Morocco

Extracts by lexical chains

- Barzilay & Elhadad'97; Silber & McCoy'02
- A chain C represents a "concept" in WordNet
  - *Financial institution* "bank"
  - *Place to sit down in the park* "bank"
  - *Sloppy land* "bank"
- A chain is a list of words, the order of the words is that of their occurrence in the text
- A noun N is inserted in C if N is related to C
  - relations used=identity; synonym; hypernym
- Compute lexical chains; score lexical chains in function of their members; select sentences according to membership to lexical chains of words in sentence

44

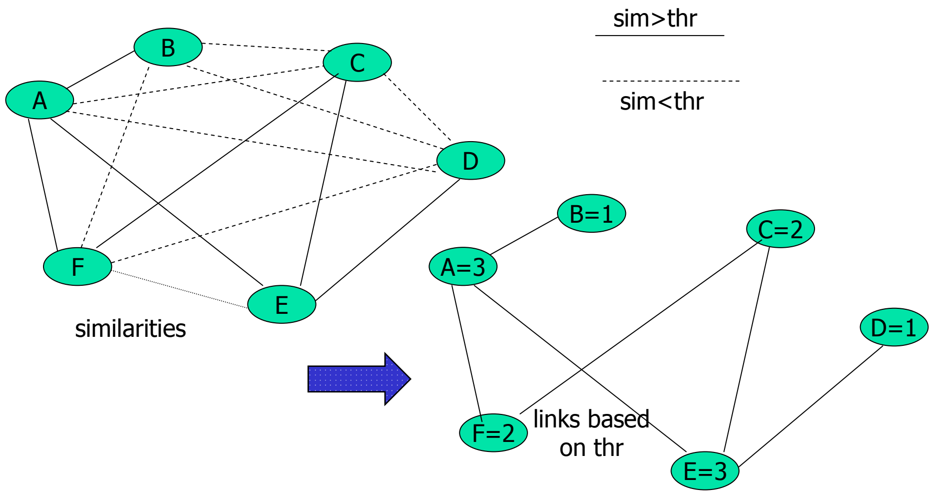
Information retrieval techniques (Salton&al'97)

- Vector Space Model
  - each text unit represented as $D_i = (d_{i1}, \ldots, d_{in})$
- Similarity metric

$$sim(D_i, D_j) = \sum d_{ik} . d_{jk}$$

- metric normalised to obtain 0-1 values
- Construct a graph of paragraphs. Strength of link is the similarity metric
- Use threshold (thr) to decide upon similar paragraphs

Text relation map



Information retrieval techniques

- identify regions where paragraphs are well connected
- paragraph selection heuristics
  - bushy path
    - select paragraphs with many connections with other paragraphs and present them in text order
  - depth-first path
    - select one paragraph with many connections; select a connected paragraph (in text order) which is also well connected; continue
  - segmented bushy path
    - follow the bushy path strategy but locally including paragraphs from all "segments of text": a bushy path is created for each segment

47

Information retrieval techniques

- Co-selection evaluation
  - because of low agreement across human annotators (~46%) new evaluation metrics were defined
  - optimistic scenario: select the human summary which gives best score
  - pessimistic scenario: select the human summary which gives worst score
  - union scenario: select the union of the human summaries
  - intersection scenario: select the overlap of human summaries

48

LREC 2008 – Marrakech - Morocco

Rhetorical analysis

- Rhetorical Structure Theory (RST)
  - Mann & Thompson'88
- Descriptive theory of text organization
- Relations between two text spans
  - nucleus & satellite (hypotactic)
  - nucleus & nucleus (paratactic)
  - "IR techniques have been used in text summarization. For example, X used term frequency. Y used tf*idf."

49

LREC 2008 – Marrakech - Morocco

Rhetorical analysis

- relations are deduced by judgement of the reader
- texts are represented as trees, internal nodes are relations
- text segments are the leafs of the tree
  - (1) Apples are very cheap. (2) Eat apples!!!
  - (1) is an argument in favour of (2), then we can say that (1) motivates (2)
  - (2) seems more important than (1), and coincides with (2) being the nucleus of the motivation

50

Rhetorical analysis

- Relations can be marked on the syntax
  - John went to sleep because he was tired.
  - Mary went to the cinema and Julie went to the theatre.
- RST authors say that markers are not necessary to identify a relation
- However all RTS analysers rely on markers
  - "however", "therefore", "and", "as a consequence", etc.
- strategy to obtain a complete tree
  - apply rhetorical parsing to "segments" (or paragraphs)
  - apply a cohesion measure (vocabulary overlap) to identify how to connect individual trees

Rhetorical analysis based summarization

(A) Smart cards are becoming more attractive

(B) as the price of micro-computing power and storage continues to drop.

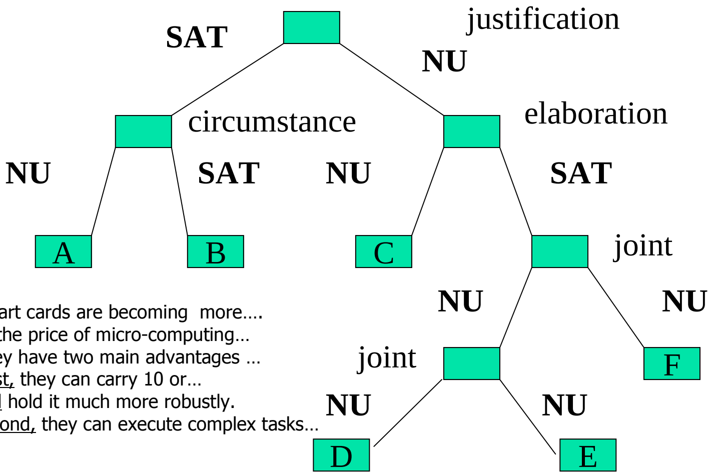
(C) They have two main advantages over magnetic strip cards.

(D) First, they can carry 10 or even 100 times as much information

(E) and hold it much more robustly.

(F) Second, they can execute complex tasks in conjunction with a terminal.

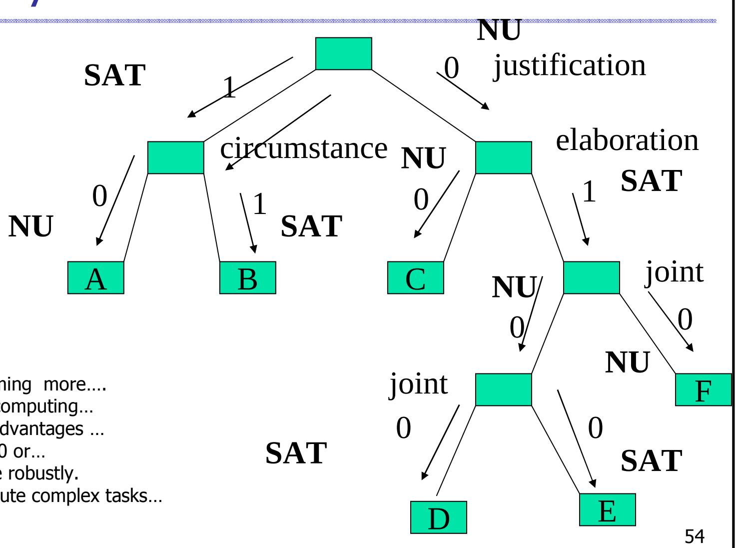
Rhetorical tree



(A) Smart cards are becoming more….
(B) as the price of micro-computing…
(C) They have two main advantages …
(D) First, they can carry 10 or…
(E) and hold it much more robustly.
(F) Second, they can execute complex tasks…

Penalty: Ono'94

Penalty
A=1
B=2
C=0
D=1
E=1
F=1



(A) Smart cards are becoming more….
(B) as the price of micro-computing…
(C) They have two main advantages …
(D) First, they can carry 10 or…
(E) and hold it much more robustly.
(F) Second, they can execute complex tasks…

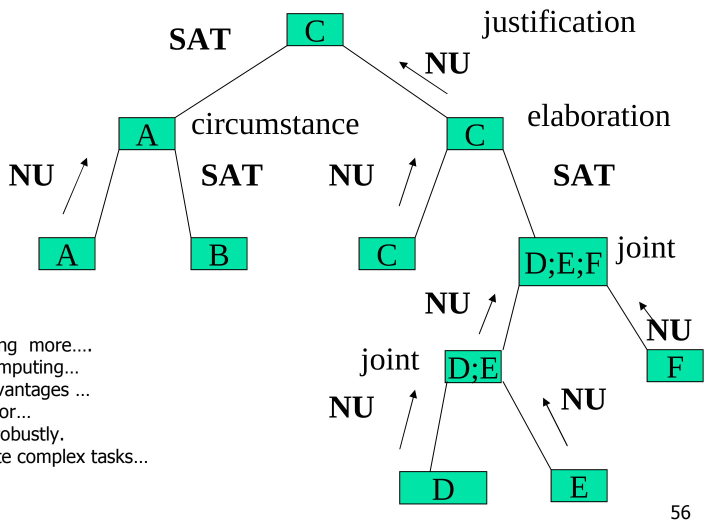
RTS extract

(C) They have two main advantages over magnetic strip cards.

(A) Smart cards are becoming more attractive
(C) They have two main advantages over magnetic strip cards.
(D) First, they can carry 10 or even 100 times as much information
(E) and hold it much more robustly.
(F) Second, they can execute complex tasks in conjunction with a terminal.

(A) Smart cards are becoming more attractive
(B) as the price of micro-computing power and storage continues to drop.
(C) They have two main advantages over magnetic strip cards.
(D) First, they can carry 10 or even 100 times as much information
(E) and hold it much more robustly.
(F) Second, they can execute complex tasks in conjunction with a terminal.

Promotion: Marcu'97



(A) Smart cards are becoming more....
(B) as the price of micro-computing...
(C) They have two main advantages ...
(D) First, they can carry 10 or...
(E) and hold it much more robustly.
(F) Second, they can execute complex tasks...

RST extract

(C) They have two main advantages over magnetic strip cards.

(A) Smart cards are becoming more attractive
(C) They have two main advantages over magnetic strip cards.

(A) Smart cards are becoming more attractive
(B) as the price of micro-computing power and storage continues to drop.
(C) They have two main advantages over magnetic strip cards.
(D) First, they can carry 10 or even 100 times as much information
(E) and hold it much more robustly.
(F) Second, they can execute complex tasks in conjunction with a terminal.

Information Extraction

ALGIERS, May 22 (AFP) - At least 538 people were killed and 4,638 injured when a powerful earthquake struck northern Algeria late Wednesday, according to the latest official toll, with the number of casualties set to rise further ... The epicentre of the quake, which measured 5.2 on the Richter scale, was located at Thenia, about 60 kilometres (40 miles) east of Algiers, ...

DATE	21/05/2003
DEATH	538
INJURED	4,638
EPICENTER	Thenia, Algeria
INTENSITY	5.2, Ritcher

FRUMP (de Jong'82)

a small earthquake shook several Southern Illinois counties Monday night, the National Earthquake Information Service in Golden, Colo., reported. Spokesman Don Finley said the quake measured 3.2 on the Richter scale, "probably not enough to do any damage or cause any injuries." The quake occurred about 7:48 p.m. CST and was centered about 30 miles east of Mount Vernon, Finlay said. It was felt in Richland, Clay, Jasper, Effington, and Marion Counties.

<u>There was an earthquake in Illinois with a 3.2 Richter scale.</u>

CBA: Concept-based Abstracting (Paice&Jones'93)

- Summaries in an specific domain, for example crop husbandry, contain specific concepts.
  - SPECIES (the crop in the study)
  - CULTIVAR (variety studied)
  - HIGH-LEVEL-PROPERTY (specific property studied of the cultivar, e.g. yield, growth)
  - PEST (the pest that attacks the cultivar)
  - AGENT (chemical or biological agent applied)
  - LOCALITY (where the study was conducted)
  - TIME (years of the study)
  - SOIL (description of the soil)

CBA

- Given a document in the domain, the objective is to instantiate with "well formed strings" each of the concepts
- CBA uses patterns which implement how the concepts are expressed in texts

  "fertilized with *procymidane*" gives the pattern "fertilized with AGENT"
- Can be quite complex and involve several concepts
  - PEST is a ? pest of SPECIES

  where ? matches a sequence of input tokens

LREC 2008 – Marrakech - Morocco

CBA

- Each pattern has a weight
- Criteria for variable instantiation
  - Variable is inside pattern
  - Variable is on the edge of the pattern
- Criteria for candidate selection
  - all hypothesis' substrings are considered
    - decease of SPECIES
    - effect of ? in SPECIES
  - count repetitions and weights
  - select one substring for each semantic role

CBA

- Canned-text based generation
  ```
  this paper studies the effect of [AGENT] on the
  [HLP] of [SPECIES] OR  this paper studies the
  effect of [METHOD] on the [HLP] of [SPECIES]
  when it is infested by [PEST]…
  ```
  Summary: *This paper studies the effect of G. pallida on the yield of potato. An experiment in 1985 and 1986 at York was undertaken.*
- evaluation
  - central and peripheral concepts
  - form of selected strings
- pattern acquisition can be done automatically
- informative summaries include verbatim "conclusive" sentences from document

Headline generation: Banko&al'00

- Generate a summary shorter than a sentence
  - Text: Acclaimed Spanish soprano de los Angeles dies in Madrid after a long illness.
  - Summary: de Los Angeles died
- Generate a sentence with pieces combined from different parts of the texts
  - Text: Spanish soprano de los Angeles dies. She was 81.
  - Summary: de Los Angeles dies at 81
- Method borrowed from statistical machine translation
  - model of word selection from the source
  - model of realization in the target language

Headline generation

- Content selection
  - how many and what words to select from document
- Content realization
  - how to put words in the appropriate sequence in the headline such that it looks ok
- training: available texts + headlines

Example

President Clinton met with his top Mideast adviser, including Secretary of State Madeleine Albright and U.S. peace envoy Dennis Ross, in preparation for a session with Isralel Prime Minister Benjamin Netanyahu tomorrow. Palestinian leader Yasser Arafat is to meet with Clinton later this week. Published reports in Israel say Netanyahu will warn Clinton that Israel can't withdraw from more than nine percent of the West Bank in its next schedulled pullback, although Clinton wants 12-15 percent pullback.

- original title: *U.S. pushes for mideast peace*
- automatic title
  - *clinton*
  - *clinton wants*
  - *clinton netanyahu arafat*
  - *clinton to mideast peace*

Cut & Paste summarization

- Cut&Paste Summarization: Jing&McKeown'00
  - "HMM" for word alignment to answer the question: what document positions a word in the summary comes from?
  - a word in a summary sentence may come from different positions, not all of them are equally likely
  - given words $I_1 \ldots I_n$ (in a summary sentence) the following probability table is needed: $P(I_{k+1}=<S2,W2>| I_k=<S1,W1>)$
  - they associate probabilities by hand following a number of heuristics
  - given a sentence summary, the alignment is computed using the Viterbi algorithm

Summary sentence:
(F0:S1 **arthur b sackler vice president for law and public policy of time warner inc**) (F1:S-1 *and*) (F2:S0 **a member of the direct marketing association told**) (F3:S2 **the communications subcommittee of the senate commerce committee**) (F4:S-1 *that legislation*) (F5:S1**to protect**) (F6:S4 **children' s**) (F7:S4 **privacy**) (F8:S4 **online**) (F9:S0 **could destroy the spontaneous nature that makes the internet unique**)

Source document sentences:
Sentence 0: a proposed new law that would require web publishers to obtain parental consent before collecting personal information from children (**F9 could destroy the spontaneous nature that makes the internet unique**) (**F2 a member of the direct marketing association told**) a senate panel thursday
Sentence 1: (**F0 arthur b sackler vice president for law and public policy of time warner inc**) said the association supported efforts (**F5 to protect**) children online but he urged lawmakers to find some middle ground that also allows for interactivity on the internet
Sentence 2: for example a child's e-mail address is necessary in order to respond to inquiries such as updates on mark mcguire's and sammy sosa's home run figures this year or updates of an online magazine sackler said in testimony to (**F3 the communications subcommittee of the senate commerce committee**)
Sentence 4: the subcommittee is considering the (**F6 children's**) (**F8 online**) (**F7 privacy**) protection act which was drafted on the recommendation of the federal trade commission

Cut & Paste

- Cut&Paste Summarization
  - Sentence reduction
    - a number of resources are used (lexicon, parser, etc.)
    - exploits connectivity of words in the document (each word is weighted)
    - uses a table of probabilities to decide when to remove a sentence component
    - final decision is based on probabilities, mandatory status, and local context
  - Rules for sentence combination were manually developed

LREC 2008 – Marrakech - Morocco

Sentence condensation

- Sentence condensation: Knight&Marcu'00
  - probabilistic framework: noisy-channel model
  - corpus: automatically collected <sentences, compressions>
  - model explains how short sentences can be re-written
  - a long sentence L can be generated from a short sentence S, two probabilities are needed
  - P(L/S) and P(S)
  - the model seeks to maximize P(L/S)xP(S)

Paraphrase

- Alignment based paraphrase: Barzilay&Lee'2003
- unsupervised approach to learn:
  - patterns in the data & equivalences among patterns
  - X injured Y people, Z seriously = Y were injured by X among them Z were in serious condition
  - learning is done over two different corpus which are comparable in content
- use a sentence clustering algorithm to group together sentences that describe similar events

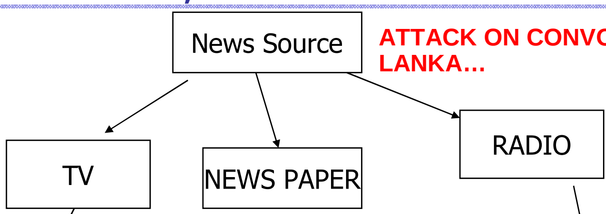
Similar event descriptions

- **Cluster of similar sentences**
  - **A Palestinian suicide bomber blew himself up in** a southern city Wednesday, **killing** two other **people and wounding** 27.
  - **A suicide bomber blew himself up in** the settlement of Efrat, on Sunday, **killing** himself **and injuring** seven people.
  - **A suicide bomber blew himself up in** the coastal resort of Netanya on Monday, **killing** three other **people and wounding** dozens more.
- **Variable substitution**
  - **A Palestinian suicide bomber blew himself up in** a southern city DATE, **killing** NUM other **people and wounding** NUM.
  - **A suicide bomber blew himself up in** the settlement of NAME, on DATE, **killing** himself **and injuring** NUM people.
  - **A suicide bomber blew himself up in** the coastal resort of NAME on NAME, **killing** NUM other **people and wounding** dozens more.

Multi-document Summarization

- Input is a set of related documents, redundancy must be avoided
- The relation can be one of the following:
  - report information on the same event or entity (e.g. documents "about" Angelina Jolie)
  - contain information on a given topic (e.g. the Iran – US relations)
  - ...

Same event, different accounts



ATTACK ON CONVOY IN SRI LANKA...

At least 13 sailors have been killed in a mine attack on a convoy in north-western Sri Lanka, officials say.

Tamil Tiger guerrillas have blown up a navy bus in northeastern Sri Lanka, killing at least 10 sailors and wounding 17 others.

Blasts blamed on Tamil Tiger rebels killed 13 people on Wednesday in Sri Lanka's northeast and dozens more were injured, officials said, raising fears planned peace talks may be cancelled and a civil war could restart.

Multi-document summarization

- Redundancy of information
  - the destruction of Rome by the Barbarians in 410....
  - Rome was destroyed by Barbarians.
  - Barbarians destroyed Rome in the V Century
  - In 410, Rome was destroyed. The Barbarians were responsible.
- fragmentary information
  - D1="earthquake in Turkey"; D2="measured 6.5"
- contradictory information
  - D1="killed 3"; D2= "killed 4"
- relations between documents
  - inter-document-coreference
  - D1="Tony Blair visited Bush"; D2="UK Prime Minister visited Bush"

LREC 2008 – Marrakech - Morocco

Similarity metrics

- text fragments (sentences, paragraphs, etc.) represented in a vector space model OR as bags of words and use set operations to compare them
- can be "normalized" (stemming, lemmatised, etc)
- stop words can be removed
- weights can be term frequencies or tf*idf…

$$D_i = (d_{i1}, \ldots, d_{in})$$

$$sim(D_i, D_j) = \sum d_{ik} . d_{jk} \qquad \cos(D_i, D_j) = \frac{\sum_k (d_{ik} . d_{jk})}{\sqrt{\sum_k (d_{ik})^2 \sum_k (d_{jk})^2}}$$

Morphological techniques

- IR techniques: a query is the input to the system
- Goldstein&al'00. Maximal Marginal Relevance
  - a formula is used allowing the inclusion of sentences relevant to the query but different from those already in the summary

similarity to query

$Q = \text{query}$
$R = \text{list of documents}$
$D_k = \text{k - document in list}$
$S = \text{subset of R already scanned}$

$$MMR(Q,R,S) = \arg\max_{D_i \in R \setminus S}(\lambda sim_1(D_i,Q) +$$
$$(\lambda - 1)\max_{D_j \in S} sim_2(D_i,D_j))$$

similarity to document already seen

Centroid-based summarization (Radev&al'00;Saggion&Gaizauskas'04)

- given a set of documents create a centroid of the cluster
  - centroid = set of words in the cluster considered "statistically" significant
  - centroid is a set of terms and weights
- centroid score = similarity between a sentence and the centroid
- combine the centroid score with document features such as position
- detect and eliminate sentence redundancy using a similarity metric

Sentence ordering

- simplest strategy is to present sentences in temporal order when date of document is known
- important for both single and multi-document summarization (Barzilay, Elhadad, McKeown'02)
- some strategies
  - Majority order
  - Chronological order
  - Combination
- probabilistic model (Lapata'03)
  - the model learns order constraints in a particular domain
  - the main component is a probability table
    - $P(S_i|S_{i-1})$ for sentences S
    - the representation of each sentence is a set of features for
      - verbs, nouns, and dependencies

Semantic techniques

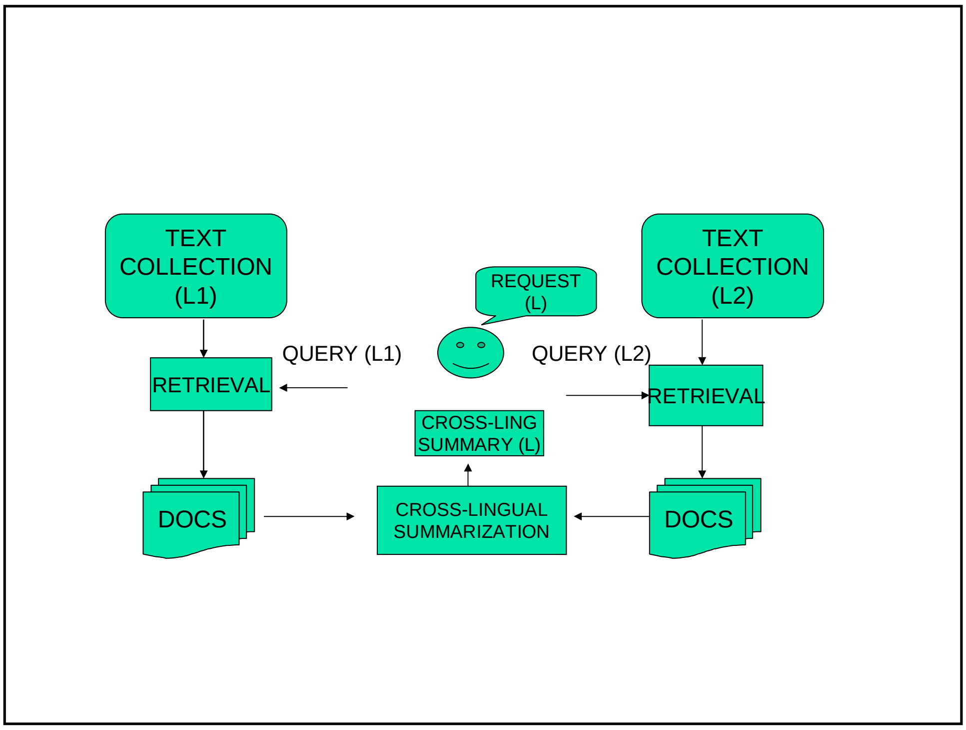
- Knowledge-based summarization in SUMMONS (Radev & McKeown'98)
- Conceptual summarization
  - reduction of content
- Linguistic summarization
  - Conciseness
- corpus of summaries
  - strategies for content selection
  - summarization lexicon
- summarization from a template knowledge base
- planning operators for content selection
  - 8 operators
- linguistic generation
  - generating summarization phrases
  - generating descriptions

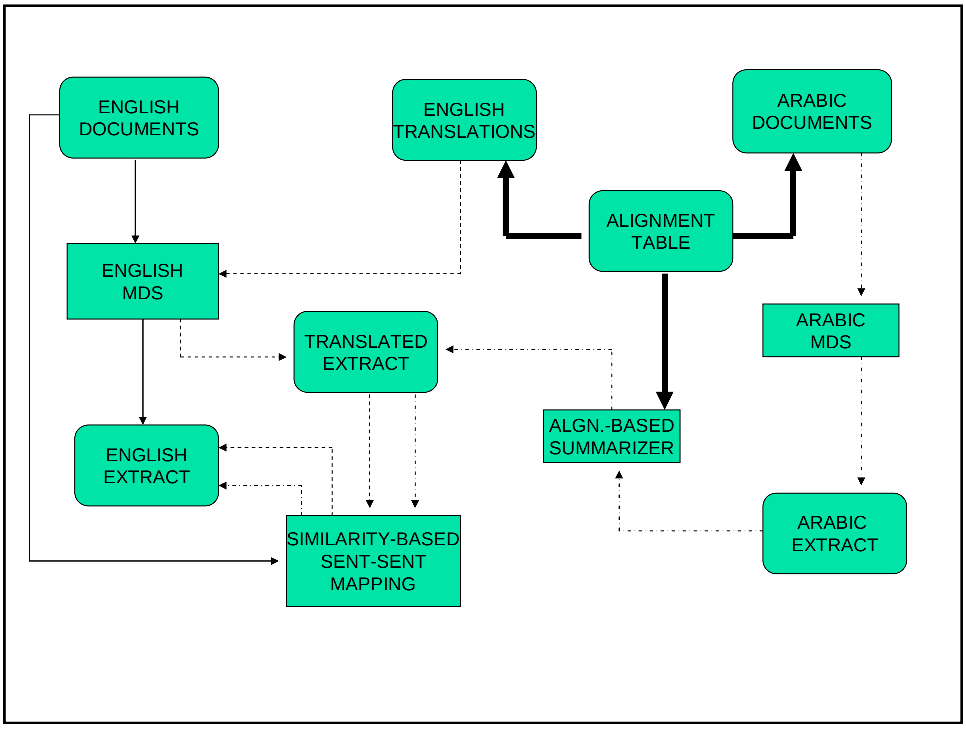
Example summary

Reuters reported that 18 people were killed on *Sunday* in a bombing in Jerusalem. *The next day*, a bomb in Tel Aviv killed at least 10 people and wounded 30 according to Israel radio. Reuters reported that *at least 12 people* were killed and *105* wounded *in the second incident*. *Later the same day*, Reuters reported that Hamas has claimed responsibility for the act.

Cross-lingual Summarization

- Given a document in language S, produce a summary of the document in language T
- Given a set of documents in languages you don't know produce a summary in a language you know
- The problem has been addressed as part of the Multilingual Summarization Evaluation (MSE) 2005-2006 but also as part of the Document Understanding Conferences
- This is a common activity – abstracts in English of documents in a language other than English have to be produced to be included in abstracting databases


TEXT COLLECTION (L1)
TEXT COLLECTION (L2)
REQUEST (L)
QUERY (L1)
QUERY (L2)
RETRIEVAL
RETRIEVAL
CROSS-LING SUMMARY (L)
DOCS
CROSS-LINGUAL SUMMARIZATION
DOCS


ENGLISH DOCUMENTS
ENGLISH TRANSLATIONS
ARABIC DOCUMENTS
ALIGNMENT TABLE
ENGLISH MDS
ARABIC MDS
TRANSLATED EXTRACT
ALGN.-BASED SUMMARIZER
ENGLISH EXTRACT
ARABIC EXTRACT
SIMILARITY-BASED SENT-SENT MAPPING

Text Summarization Evaluation

- Identify when a particular algorithm can be used commercially
- Identify the contribution of a system component to the overall performance
- Adjust system parameters
- Objective framework to compare own work with work of colleagues
- Expensive because requires the construction of standard sets of data and evaluation metrics
- May involve human judgement
- There is disagreement among judges
- Automatic evaluation would be ideal but not always possible

Intrinsic Evaluation

- Summary evaluated on its own or comparing it with the source
  - Is the text cohesive and coherent?
  - Does it contain the main topics of the document?
  - Are important topics omitted?
  - Compare summary with ideal summaries

How intrinsic evaluation works with ideal summaries?

- Given a machine summary (P) compare to one or more human summaries (M) using a scoring function score(P,M), aggregate the scores per system, use the aggregated score to rank systems
- Compute confidence values to detect true system differences (e.g. score(A) > score(B) does not guarantee A better than B)

Extrinsic Evaluation

- Evaluation in an specific task
  - Can the summary be used instead of the document?
    - Can the document be classified by reading the summary?
    - Can we answer questions by reading the summary?

Evaluation of extracts

	System	
Human	+	-
+	TP	FN
-	FP	TN

- precision (P) $\frac{TP}{TP+FP}$
- recall (R) $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$
- F-score (F) $\frac{(\beta^2+1)P.R}{\beta^2 P+R}$
- Accuracy (A) $\frac{TP+TN}{TP+FP+FP+FN}$

LREC 2008 – Marrakech - Morocco

Evaluation of extracts

- Relative utility (fuzzy) (Radev&al'00)
  - each sentence has a degree of "belonging to a summary"
  - H={(S1,10), (S2,7),...(Sn,1)}
  - A={ S2,S5,Sn } => val(S2) + val(S5) + val(Sn)
  - Normalize dividing by maximum

Other metrics

- Content based metrics
  - the fragments bellow are similar, however for precision and recall do not count as such
    - "three people were killed in the blast" vs "In the blast, 3 were killed"
  - overlap
    - Based on set n-gram intersection
    - Fine grained metrics than combine different sets of n-grams can be used
  - cosine in Vector Space Model
  - Longest subsequence
    - Minimal number of deletions/insertions needed to obtain two identical chains
  - Do they really measure semantic content?

LREC 2008 – Marrakech - Morocco

SUMMAC evaluation

- High scale system independent evaluation
- basically extrinsic
- 16 systems
- summaries in tasks carried out by defence analysis of the American government

SUMMAC tasks

- "ad hoc" task
  - indicative summaries
  - system receives a document + a topic and has to produce a topic-based
  - analyst has to classify the document in two categories
    - Document deals with topic
    - Document does not deal with topic

LREC 2008 – Marrakech - Morocco

SUMMAC tasks

- Categorization task
  - generic summaries
  - given n categories and a summary, the analyst has to classify the document in one of the n categories or none of them
  - one wants to measure whether summaries reduce classification time without loosing classification accuracy

SUMMAC experiments

- Experimental conditions
  - text: full-document; fixed-length summary; variable-length summary; default summary (baseline)
  - technology: each of the participants
  - consistency: 51 analysts

LREC 2008 – Marrakech - Morocco

SUMMAC

- data
  - “ad hoc”: 20 topics each with 50 documents
  - categorization: 10 topics each with 100 documents (5 categories)
- Results “ad hoc” task
  - Variable length summaries take less time to classify by a factor of 2 (33.12 sec/doc vs. 58.89 sec/doc with full-text)
  - Classification accuracy reduced but <u>not significantly</u>

SUMMAC

- Results of categorization task
  - only significant differences in time between 10% length summaries and full-documents
  - no difference in classification accuracy
  - many FN observed (automatic summaries lack many relevant topics)
- 3 groups of systems observed
- ad hoc: pair-wise human agreement 69%; 53% 3-way; 16% unanimous

DUC experience

- National Institute of Standards and Technology (NIST)
- further progress in summarization and enable researchers participate in large-scale experiments
- Document Understanding Conference
  - 2000-2006
  - from 2008 Text Analysis Conference (TAC)

DUC 2004

- Tasks for 2004
  - Task 1: very short summary
  - Task 2: short summary of cluster of documents
  - Task 3: very short cross-lingual summary
  - Task 4: short cross-lingual summary of document cluster
  - Task 5: short person profile
- Very short (VS) summary <= 75 bytes
- Short (S) summary <= 665 bytes

LREC 2008 – Marrakech - Morocco

DUC 2004 - Data

- 50 TDT English news clusters (tasks 1 & 2) from AP and NYT sources
  - 10 docs/topic
  - Manual S and VS summaries
- 24 TDT Arabic news clusters (tasks 3 & 4) from France Press
  - 13 topics as before and 12 new topics
  - 10 docs/topic
  - Related English documents available
  - IBM and ISI machine translation systems
  - S and VS summaries created from manual translations
- 50 TREC English news clusters from NYT, AP, XIE
  - Each cluster with documents which contribute to answering “Who is X?”
  - 10 docs/topic
  - Manual S summaries created

DUC 2004 - Tasks

- Task 1
  - VS summary of each document in a cluster
  - Baseline = first 75 bytes of document
  - Evaluation = ROUGE
- Task 2
  - S summary of a document cluster
  - Baseline = first 665 bytes of most recent document
  - Evaluation = ROUGE

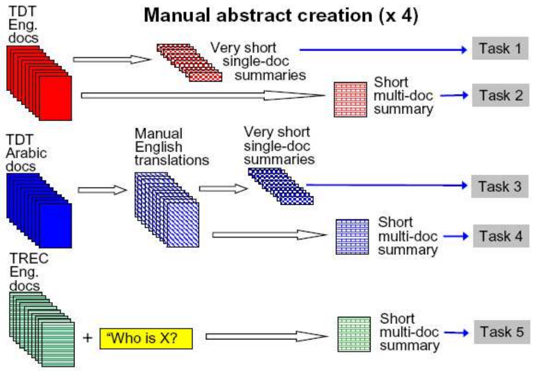
DUC 2004 - Tasks

- Task 3
  - VS summary of each translated document
  - Use: automatic translations; manual translations; automatic translations + related English documents
  - Baseline = first 75 bytes of best translation
  - Evaluation = ROUGE
- Task 4
  - S summary of a document cluster
  - Use: same as for task 3
  - Baseline = first 665 bytes of most recent best translated document
  - Evaluation = ROUGE
- Task 5
  - S summary of document cluster + "Who is X?"
  - Evaluation = using Summary Evaluation Environment (SEE): quality & coverage; ROUGE

LREC 2008 – Marrakech - Morocco

SLIDE FROM Document Understanding Conferences

Summary of tasks



103

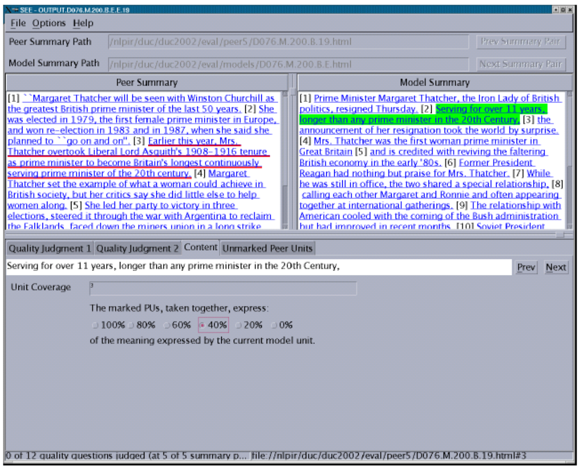
LREC 2008 – Marrakech - Morocco

DUC 2004 – Human Evaluation

- Human summaries segmented in Model Units (MUs)
- Submitted summaries segmented in Peer Units (PUs)
- For each MU
  - Mark all PUs sharing content with the MU
  - Indicates whether the Pus express 0%, 20%,40%,60%,80%,100% of MU
  - For all non-marked PU indicate whether 0%,20%,...100% of PUs are related but needn't to be in summary

104

Summary evaluation environment (SEE)



DUC 2004 – Questions

- 7 quality questions
- 1) Does the summary build from sentence to sentence to a coherent body of information about the topic?
  - A. Very coherently
  - B. Somewhat coherently
  - C. Neutral as to coherence
  - D. Not so coherently
  - E. Incoherent
- 2) If you were editing the summary to make it more concise and to the point, how much useless, confusing or repetitive text would you remove from the existing summary?
  - A. None
  - B. A little
  - C. Some
  - D. A lot
  - E. Most of the text

DUC 2004 - Questions

- Read summary and answer the question
- Responsiveness (Task 5)
  - Given a question "Who is X" and a summary
  - Grade the summary according to how responsive it is to the question
    - 0 (worst) - 4 (best)

LREC 2008 – Marrakech - Morocco

ROUGE package

- Recall-Oriented Understudy for Gisting Evaluation
- Developed by Chin-Yew Lin at ISI (see DUC 2004 paper)
- Measures quality of a summary by comparison with ideal(s) summaries
- Metrics count the number of overlapping units

ROUGE package

- ROUGE-N: N-gram co-occurrence statistics is a recall oriented metric

S1- Police killed the gunman

S2- Police kill the gunman

S3- The gunman kill police

S2=S3

ROUGE Formula

$$\mathrm{ROUGE-n} = \frac{\sum_{S \in \{Refs\}} \sum_{n\text{-}gram \in S} count_{match}(n\text{-}gram)}{\sum_{S \in \{Refs\}} \sum_{n\text{-}gram \in S} count(n\text{-}gram)}$$

LREC 2008 – Marrakech - Morocco

ROUGE package

- ROUGE-L: Based on longest common subsequence

S1- Police killed the gunman

S2- <u>Police</u> kill <u>the gunman</u>

S3- <u>The gunman</u> kill police

S2 better than S3

111

LREC 2008 – Marrakech - Morocco

Example (R-1 and R-L)

- Peer: At least 13 sailors have been killed in a mine attack on a convoy in north-western Sri Lanka, officials say.
- Model-1: Tamil Tiger guerrillas have blown up a navy bus in northeastern Sri Lanka, killing at least 10 sailors and wounding 17 others.
- Model-1: Blasts blamed on Tamil Tiger rebels killed 13 people on Wednesday in Sri Lanka's northeast and dozens more were injured, officials said, raising fears planned peace talks may be cancelled and a civil war could restart.

ROUGE-1
- Peer has 21 1-grams (x2 = 42)
- Model-1 has 22
- Model-2 has 37 (total = 59)
- 1-grams hits 16
- 1-gram recall 0.27
- 1-gram precision 0.38
- 1-gram f-score 0.31

ROUGE-L

LCS: have a in sri lanka

LCS: killed on in sri lanka officials
- Peer has 21 words (x2 = 42)
- Model-1 has 22
- Model-2 has 37 (total = 59)
- LCS-hits is 11
- LCS recall 0.18
- LCS precision 0.26
- LCS f-score 0.21

112

ROUGE package

- ROUGE-W: weighted longest common subsequence, favours consecutive matches
- ROUGE-S: Skip-bigram recall metric
- Arbitrary in-sequence bigrams are computed
- ROUGE-SU adds unigrams to ROUGE-S

ROUGE package

- Co-relation with human judgment
- Experiments on DUC 2000-2003 data
- 17 ROUGE metrics tested
- Pearson's correlation coefficients computed

ROUGE Results

- ROUGE-S4, S9, and ROUGE-W1.2 were the best in 100 words single doc task, but were statistically indistinguishable from most other ROUGE metrics.
- ROUGE-1, ROUGE-L, ROUGE-SU4, ROUGE-SU9, and ROUGE-W1.2 worked very well in 10 words headline like task (Pearson's ρ ~ 97%).
- ROUGE-1, 2, and ROUGE-SU* were the best in 100 words multi-doc task but were statistically equivalent to other ROUGE-S and SU metrics.
- ROUGE-1, 2, ROUGE-S, and SU worked well in other multi-doc tasks.

Pyramids

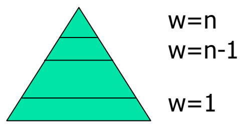
- Human evaluation of content: Nenkova & Passonneau (2004)
- based on the distribution of content in a pool of summaries
- Summarization Content Units (SCU):
  - fragments from summaries
  - identification of **similar fragments** across summaries

Pyramids

- SCU have
  - id, a weight, a NL description, and a set of contributors
- SCU1 (w=4) (all similar/identical content)
  - A1 - two Libyans indicted
  - B1 - two Libyans indicted
  - C1 - two Libyans accused
  - D2 – two Libyans suspects were indicted

Pyramids

- a "pyramid" of SCUs of height n is created for n gold standard summaries
- each SCU in tier T_i in the pyramid has weight i
- with highly weighted SCU on top of the pyramid
- the best summary is one which contains all units of level n, then all units from n-1,…
- if D_i is the number of SCU in a summary which appear in T_i for summary D, then the weight of the summary is:



$$D = \sum_{i=1}^{n} i * D_i$$

Pyramids score

- let X be the total number of units in a summary
- it is shown that more than 4 ideal summaries are required to produce reliable rankings

$$Max = \sum_{i=j+1}^{n} i * |T_i| + j * (X - \sum_{i=j+1}^{n} |T_i|)$$

$$j = \max_{i} (\sum_{t=i}^{n} |T_t| \geq X)$$

$$Score = D / Max$$

DUC 2005

- Topic based summarization
  - given a set of documents and a topic description, generate a 250 words summary

```
<TOPIC ID="d324e" GRANULARITY="specific">
How have relations between Argentina and Great Britain developed since the 1982 war over the Falkland Islands? Have
diplomatic, economic, and military relations been restored? Do differences remain over the status of the Falkland Islands?
</TOPIC>

<TOPIC ID="d332h" GRANULARITY="general">
What kinds of non-tax crimes have lead to tax evasion prosecutions (failure to file, inaccurate filing), instead of or in addition
to prosecution for the non-tax crimes themselves?
</TOPIC>
```

- Evaluation
  - ROUGE
  - Pyramid

Other evaluations

- Multilingual Summarization Evaluation (MSE) 2005 and 2006
  - basically task 4 of DUC 2004
  - Arabic/English multi-document summarization
  - human evaluation with pyramids
  - automatic evaluation with ROUGE

LREC 2008 – Marrakech - Morocco

Other evaluations

- Text Summarization Challenge (TSC)
  - Summarization in Japan
  - Two tasks in TSC-2
    - A: generic single document summarization
    - B: topic based multi-document summarization
  - Evaluation
    - summaries ranked by content & readability
    - summaries scored in function of a revision based evaluation metric
- Text Analysis Conference 2008 (http://www.nist.gov/tac)
  - Summarization, QA, Textual Entailment

MEAD

- Dragomir Radev and others at University of Michigan
- publicly available toolkit for multi-lingual summarization and evaluation
- implements different algorithms: position-based, centroid-based, it*idf, query-based summarization
- implements evaluation methods: co-selection, relative-utility, content-based metrics

MEAD

- Perl & XML-related Perl modules
- runs on POSIX-conforming operating systems
- English and Chinese
- summarizes single documents and clusters of documents

MEAD

- compression = words or sentences; percent or absolute
- output = console or specific file
- ready-made summarizers
  - lead-based
  - random

MEAD architecture

- configuration files
- feature computation scripts
- classifiers
- re-rankers

Configuration file

```
<MEAD-CONFIG TARGET='GA3' LANG='ENG' CLUSTER-PATH='/clair4/mead/data/GA3'
  DATA-DIRECTORY='/clair4/mead/data/GA3/docsent'>

<FEATURE-SET BASE-DIRECTORY='/clair4/mead/data/GA3/feature/'>
  <FEATURE NAME='Centroid'
SCRIPT='/clair4/mead/bin/feature-scripts/Centroid.pl HK-WORD-enidf ENG'/>
  <FEATURE NAME='Position'
SCRIPT='/clair4/mead/bin/feature-scripts/Position.pl'/>
  <FEATURE NAME='Length'
SCRIPT='/clair4/mead/bin/feature-scripts/Length.pl'/>
</FEATURE-SET>

<CLASSIFIER COMMAND-LINE='/clair4/mead/bin/defalut-classifier.pl \
  Centroid 1 Position 1 Length 9' SYSTEM='MEADORIG' RUN='10/09'/>

<RERANKER COMMAND-LINE='/clair4/mead/bin/default-reranker.pl MEAD-cosine 0.7'/>

<COMPRESSION BASIS='sentences' PERCENT='20'/>

</MEAD-CONFIG>
```


clusters & sentences

```
<?xml version='1.0'?>
<!DOCTYPE CLUSTER SYSTEM '/clair4/mead/dtd/cluster.dtd'>

<CLUSTER LANG='ENG'>
        <D DID='41' />
        <D DID='81' />
        <D DID='87' />
</CLUSTER>
```

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE DOCSENT SYSTEM '/clair4/mead/dtd/docsent.dtd'>

<DOCSENT DID='41' LANG='ENG'>
<BODY>
<HEADLINE>
<S PAR="1" RSNT="1" SNO="1">Egyptians Suffer Second Air
Tragedy in a Year </S>
</HEADLINE>
<TEXT>
<S PAR='2' RSNT='1' SNO='2'>CAIRO, Egypt   -- The crash of a
Gulf Air flight that killed 143 people in Bahrain is a disturbing
deja vu for Egyptians: It is the second plane crash within a
year to devastate this Arab country.</S>
<S PAR='2' RSNT='2' SNO='3'>Sixty-three Egyptians were on
board the Airbus A320, which crashed into shallow Persian Gulf
waters Wednesday night after circling and trying to land in
Bahrain.</S>
```

extract & summary

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE EXTRACT SYSTEM '/clair/tools/mead/dtd/extract.dtd'>

<EXTRACT QID='GA3' LANG='ENG' COMPRESSION='7'
SYSTEM='MEADORIG' RUN='Sun Oct 13 11:01:19 2002'>
<S ORDER='1' DID='41' SNO='2' />
<S ORDER='2' DID='41' SNO='3' />
<S ORDER='3' DID='41' SNO='11' />
<S ORDER='4' DID='81' SNO='3' />
<S ORDER='5' DID='81' SNO='7' />
<S ORDER='6' DID='87' SNO='2' />
<S ORDER='7' DID='87' SNO='3' />
</EXTRACT>
```

```
[1]The Disaster Relief Fund Advisory Committee has approved a
grant of $3 million to Hong Kong Red Cross for emergency relief
for flood victims in Jiangxi, Hunan and Hubei, the Mainland.
[2]Together with the earlier grant of $3 million to World Vision
Hong Kong, the Advisory Committee has so far approved $6 million from the
Disaster Relief Fund for relief projects to assist the victims
affected by the recent floods in the Mainland.
```


LREC 2008 – Marrakech - Morocco

Mead at work

- Mead computes sentence features (real-valued)
  - position, length, centroid, etc.
  - similarity with first, is longest sentence, various query-based features
- Mead combines features
- Mead re-rank sentences to avoid repetition

Summarization with GATE - SUMMA

- GATE (http://gate.ac.uk)
  - General Architecture for Text Engineering
  - Processing & Language Resources
  - Documents follow the TIPTSTER architecture
- Text Summarization in GATE - SUMMA
  - processing resources compute feature-values for each sentence in a document
  - features are stored in documents
  - feature-values are combined to score sentences
  - need gate + summarization jar file + creole.xml

131

GATE (Cunningham&al'02)

- Framework for development and deployment of natural language processing applications
- A graphical user interface allows users (computational linguists) access, composition and visualisation of different components and experimentation
- A Java library (gate.jar) for programmers to implement and pack applications

132

Component Model

- Language Resources (LR)
  - data
- Processing Resources (PR)
  - algorithms
- Visualisation Resources (VR)
  - graphical user interfaces (GUI)

- Components are extendable and user-customisable
  - for example adaptation of an information extraction application to a new domain
  - to a new language where the change involves adaptation of a module for word recognition and sentence recognition

Documents in GATE

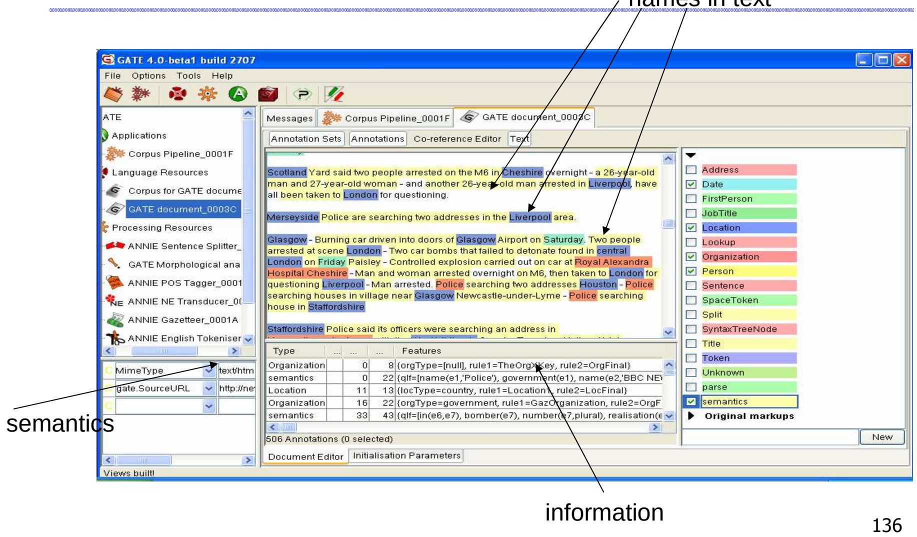
- A document is created from a file located somewhere in your disk or in a remote place or from a string
- A GATE document contains the “text” of your file and sets of annotations
- When the document is created and if a format analyser for your type is available “parsing” (format) will be applied and annotations will be created
  - xml, sgml, html, etc.
- Documents also store features, useful for representing metadata about the document
  - some features are created by GATE

Documents in GATE

- Annotations have
  - types (e.g. Token)
  - belong to particular annotation sets
  - start and end offsets – where in the document
  - features and values which are used to store orthographic, grammatical, semantic information, etc.
- Documents can be grouped in a Corpus
- Corpus is other language resource in GATE which implements a set of documents

LREC 2008 – Marrakech - Morocco

Documents in GATE

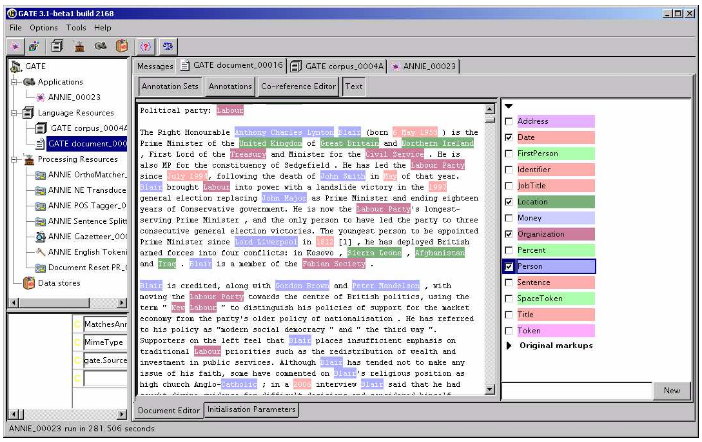


Applications in GATE

- Applications are created by sequencing processing resources
- Applications can be run over a Corpus of documents – corpus pipeline
  - so each component is applied to each document in the corpus in sequence
- Applications may not have a corpus as input, but different parameters – pipeline

137

Name Entity Recognition



138

Text Processing Tools

- Document Structure Analysis
  - different document parsers take care of the structure of your document (xml, html, etc.)
- Tokenisation
- Sentence Identification
- Parts of speech tagging
- Morphological analysis

- All these language resources have as runtime parameter a GATE document, and they will produce annotations over it
- Most resources have initialisation parameters

Summarization with GATE

- Implemented in JAVA, uses GATE documents to store information (feature, values)
- platform independent
  - Windows, Unix, Linux
- Java library which can be used to create summarization applications
- The system computes a score for each sentence and top ranked sentences are "selected" for an extract

Applications

- Single document summarization for English, Swedish, Latvian, Spanish, etc.
- Multi-document summarization for English and Arabic – centroid-based summmarization
- Cross-lingual summarization (Arabic-English)
- Profile-based summarization

LREC 2008 – Marrakech - Morocco

Resources

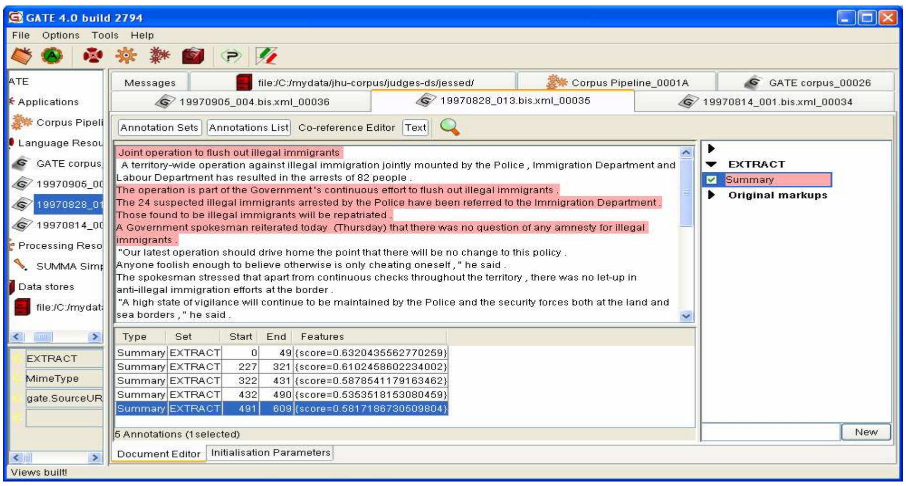
- Components to use and create IDF tables as language resources
- Vector Space Model implemented to represent text units (e.g. sentences) as vectors of terms
  - Cosine metric used to measure similarity between units
- N-gram computation and N-gram similarity computation

Feature Computation (some)

- Each feature value is numeric and it is stored as a feature of each sentence
- Position scorer (absolute, relative)
- Title scorer (similarity between sentence and title)
- Query scorer (similarity between query and sentence)
- Term Frequency scorer (sums tf*idf of sentence terms)
- Centroid scorer (similarity between a cluster centroid and a sentence – used in MDS applications)
- Features are combined using weights to produce a sentence score, this is used for sentence ranking and extraction

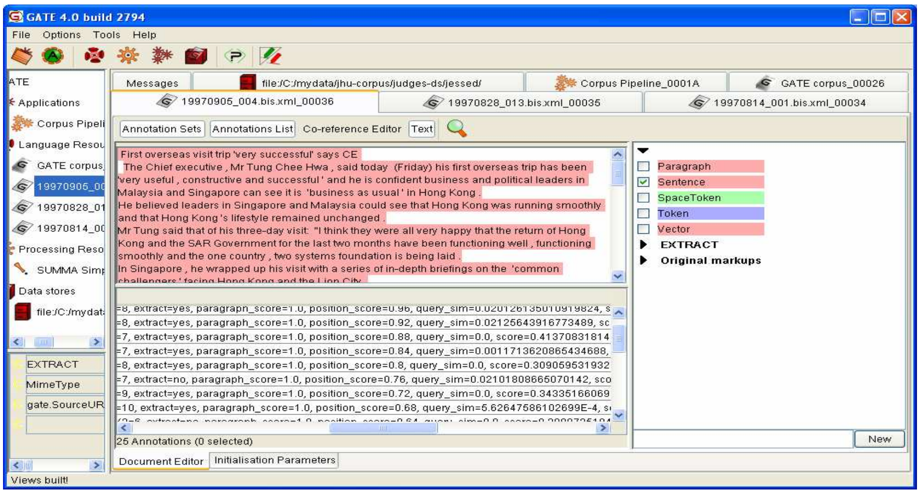
143

Sentences selected for summary



144

Features computed for each sentence



Summarizer can be trained

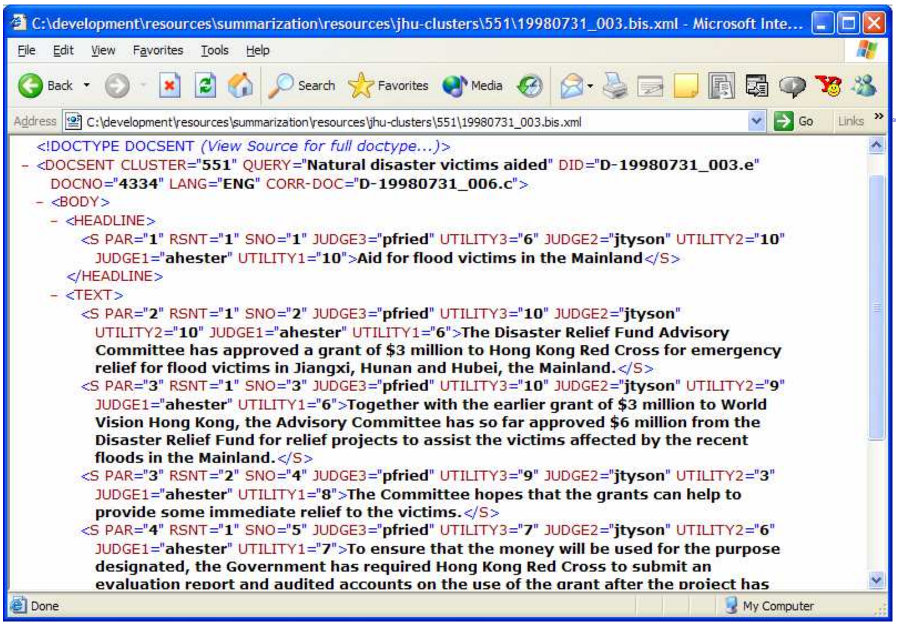
- GATE incorporates ML functionalities through WEKA (Witten&Frank'99) and LibSVM package (http://www.csie.ntu.edu.tw/~cjlin/libsvm)
- training and testing modes are available
  - annotate sentences selected by humans as keys (this can be done with a number of resources to be presented)
  - annotate sentences with feature-values
  - learn model
  - use model for creating extracts of new documents

SummBank

- Johns Hopkins Summer Workshop 2001
- Language Data Consortium (LDC)
- Drago Radev, Simone Teufel, Wai Lam, Horacio Saggion
- Development & implementation of resources for experimentation in text summarization
- http://www.summarization.com

SummBank

- Hong Kong News Corpus
- formatted in XML
- 40 topics/themes identified by LDC
- creation of a list of relevant documents for each topic
- 10 documents selected for each topic = clusters
- 3 judges evaluate each sentence in each document
- relevance judgements associated to each sentence (relative utility)
- these are values between 0-10 representing how relevant is the sentence to the theme of the cluster
- they also created multi-document summaries at different compression rates (50 words, 100 words, etc.)




```
<!DOCTYPE DOCSENT (View Source for full doctype...)>
- <DOCSENT CLUSTER="551" QUERY="Natural disaster victims aided" DID="D-19980731_003.e"
    DOCNO="4334" LANG="ENG" CORR-DOC="D-19980731_006.c">
  - <BODY>
    - <HEADLINE>
        <S PAR="1" RSNT="1" SNO="1" JUDGE3="pfried" UTILITY3="6" JUDGE2="jtyson" UTILITY2="10"
          JUDGE1="ahester" UTILITY1="10">Aid for flood victims in the Mainland</S>
      </HEADLINE>
    - <TEXT>
        <S PAR="2" RSNT="1" SNO="2" JUDGE3="pfried" UTILITY3="10" JUDGE2="jtyson"
          UTILITY2="10" JUDGE1="ahester" UTILITY1="6">The Disaster Relief Fund Advisory
          Committee has approved a grant of $3 million to Hong Kong Red Cross for emergency
          relief for flood victims in Jiangxi, Hunan and Hubei, the Mainland.</S>
        <S PAR="3" RSNT="1" SNO="3" JUDGE3="pfried" UTILITY3="10" JUDGE2="jtyson" UTILITY2="9"
          JUDGE1="ahester" UTILITY1="6">Together with the earlier grant of $3 million to World
          Vision Hong Kong, the Advisory Committee has so far approved $6 million from the
          Disaster Relief Fund for relief projects to assist the victims affected by the recent
          floods in the Mainland.</S>
        <S PAR="3" RSNT="2" SNO="4" JUDGE3="pfried" UTILITY3="9" JUDGE2="jtyson" UTILITY2="3"
          JUDGE1="ahester" UTILITY1="8">The Committee hopes that the grants can help to
          provide some immediate relief to the victims.</S>
        <S PAR="4" RSNT="1" SNO="5" JUDGE3="pfried" UTILITY3="7" JUDGE2="jtyson" UTILITY2="6"
          JUDGE1="ahester" UTILITY1="7">To ensure that the money will be used for the purpose
          designated, the Government has required Hong Kong Red Cross to submit an
          evaluation report and audited accounts on the use of the grant after the project has
```


149

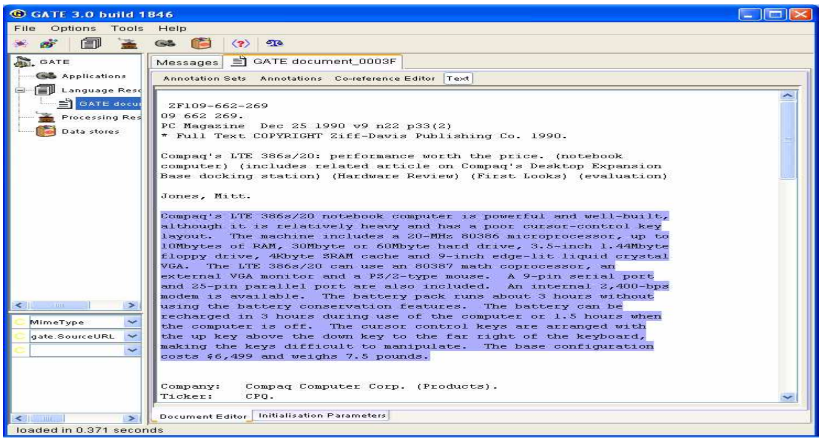
LREC 2008 – Marrakech - Morocco

Ziff-Davis Corpus for Summarization

- Each document contains the DOC, DOCNO, and TEXT fields, etc.
- The SUMMARY field contains a summary of the full text within the TEXT field.
- The TEXT has been marked with ideal extracts at the clause level.

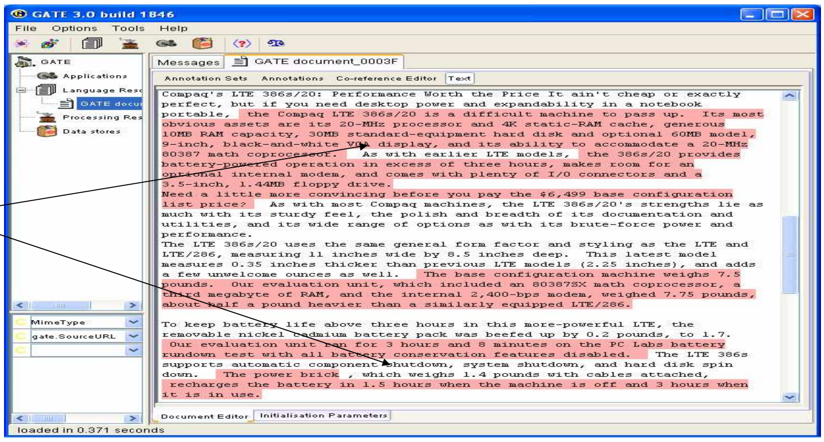
150

Document Summary



Clause Extract

clause deletion



The extracts

- Marcu’99
- Greedy-based clause rejection algorithm
  - clauses obtained by segmentation
  - “best” set of clauses
  - reject sentence such that the resulting extract is closer to the ideal summary
- Study of sentence compression
  - following Knight & Marcu’01
- Study of sentence combination
  - following Jing&McKeown’00

153

Other corpora

- SumTime-Meteo (Sripada&Reiter’05)
  - University of Aberdeen
  - (http://www.siggen.org/)
  - weather data to text
- KTH eXtract Corpus (Dalianis&Hassel’01)
  - Stockholm University and KTH
  - news articles (Swedish & Danish)
  - various sentence extracts per document

154

Other corpora

- University of Woverhampton
- CAST (Computer-Aided Summarisation Tool) Project (Hasler&Orasan&Mitkov'03)
- newswire texts + popular science
- annotated with:
  - essential sentences
  - unessential fragments in those sentences
  - links between sentences when one is needed for the understanding of the other

QA Task

- Given a question in natural language and a given text collection (or data base)
- Find the answer to the question in the collection (or data base)
- A collection can be a fixed set of documents or the Web
- Different from Information or Document retrieval which provides lists of documents matching specific queries or users' information needs

QA Task

- In the Text Retrieval Conferences (TREC) Question Answering evaluation, 3 types of questions are identified
- Factoid questions such as:
  - "Who is Tom Cruise married to?"
- List questions such as:
  - "What countries have atomic bombs?"
- Definition questions such as:
  - "Who is Aaron Copland?" or "What is aspirin?"

(Changed name to "other" question type)

QA Task

- A collection of documents is given to the participants
  - AP newswire (1998-2000), New York Times newswire (1998-2000), Xinhua News Agency (English portion, 1996-2000)
  - Approximately 1,033,000 documents and 3 gigabytes of text

QA Task

- In addition to answer the question systems have to provide a "justification" for the answer, e.g., a document where the answer occurs and which gives the possibility of fact checking
  - Who is Tom Cruise married to?
  - Nicole Kidman

  …Batman star George Clooney and Tom Cruise's wife Nicole Kidman …

QA Task

- Question can be stated in a "context-free" environment
  - "Who was Aaron Copland?"
  - "When was the South Pole reached for the first time?"
- Question may depend on previous question or answer
  - "What was Aaron Copland first ballet?"
  - "When was its premiere?"
  - "When was the South Pole reached?"
  - "Who was in charge of the expedition?"

QA Challenge

- Language variability (paraphrase)
  - Who is the President of Argentina?
  - Kirshner is the President of Argentina
  - The President of Argentina, N. Kirshner
  - N. Kirshner, the Argentinean President
  - The presidents of Argentina, N. Kirshner and Brazil, I.L da Silva…
  - Kishner is elected President of Argentina…

LREC 2008 – Marrakech - Morocco

QA Challenge

- How to locate the information given the question keywords
  - there is a gap between the wording of the question and the answer in the document collection
- Because QA is open domain it is unlikely that a system will have all necessary resources pre-computed to locate answers
  - should we have encyclopaedic knowledge in the system? all bird names, all capital cities, all drug names…
  - current systems exploit web redundancy in order to find answers, so vocabulary variation is not an issue…because of redundancy it is possible that one of the variations will exist on the Web…but what occurs in domains where information is unique…

QA Challenge

- Sometimes the task requires some deduction or extra linguistic knowledge:
  - What was the most powerful earthquake to hit Turkey?
  1. Find all earthquakes in Turkey
  2. Find intensity for each of those
  3. Pick up the one with higher intensity

  (some text-based QA systems will find the answer because it is explicitly expressed in text: "The most powerful earthquake in the history of Turkey…."

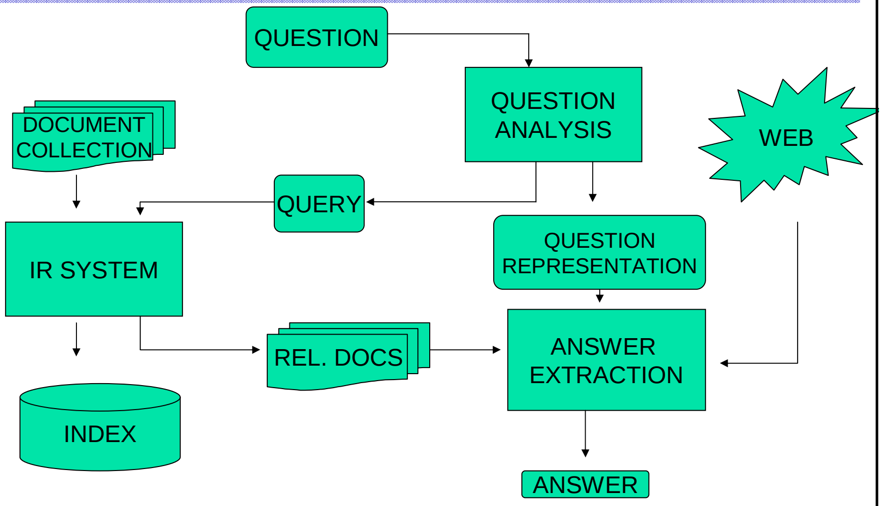
LREC 2008 – Marrakech - Morocco

How to attack the problem?

- Given a question, we could go document by document verifying if it contains the answer
- However, a more practical approach is to have the collection pre-indexed (so we know what terms belong to which document) and use a query to find a set of documents matching the question terms
- This set of matching documents is (depending on the system) further ranked to produce a list where the top document is the most likely to match the question terms
- The document ranking is generally used to inform answer extraction components

LREC 2008 – Marrakech - Morocco

QA Architecture



165

LREC 2008 – Marrakech - Morocco

Metrics and Scoring

- The principal metric for TREC8-10 was Mean Reciprocal Rank (MRR)
  - Correct answer at rank 1 scores 1
  - Correct answer at rank 2 scores 1/2
  - …

  Sum over all questions and divide by number of questions

$$MRR = \frac{\sum_{i=1}^{N} r_i}{N}$$

- where

  N = # questions, r_i = the reciprocal of the best (lowest) rank assigned by a system at which a correct answer is found for question i, or 0 if no correct answer was found
- Judgements made by human judges based on answer string alone (lenient evaluation) and by reference to documents (strict evaluation)

166

Metrics and Scoring – CWS

- The principal metric for TREC2002 was Confidence Weighted Score

$$\text{confidence weighted score} = \frac{\sum_{i=1}^{Q} \#\text{correct in first } i \text{ positions}/i}{Q}$$

- where Q is number of questions
- When only one answer is accepted per question, the metric used is answer accuracy: percent of correct answers

Answering Definition Questions

- text collection (e.g., AQUAINT)
- definition question (e.g., "*What is Goth*?", "*Who is Aaron Copland?")*
  - Goth is the definiendum or term to be defined
- answer for Goth: "*a subculture that started as one component of the punk rock scene*" or "*horror/mystery literature that is dark, eerie, and gloomy*" or ...
- architecture: Information Retrieval + Information Extraction
- <u>definiendum</u> gives little information for retrieving definition-bearing passages

Gold standard by NIST

Qid 1901: Who is Aaron Copland?

1901 1	vital	american composer
1901 2	vital	musical achievements ballets symphonies
1901 3	vital	born brooklyn ny 1900
1901 4	okay	son jewish immigrant
1901 5	okay	american communist
1901 6	okay	civil rights advocate
1901 7	okay	had senile dementia
1901 8	vital	established home for composers
1901 9	okay	won oscar for "the Heiress"
1901 10	okay	homosexual
1901 11	okay	teacher tanglewood music center boston symphony

LREC 2008 – Marrakech - Morocco

BBN Approach (Yang et al'03) – best approach in TREC 2003

1. Identify type of question (who or what) and the question target
2. Retrieve 1000 documents using an IR system and the target as query
3. For each sentence in the documents decide if it mention the target
4. Extract *kernel facts* (phrases) from each sentence
5. Rank all kernel facts according to type and similarity to a question profile (centroid)
6. Detect redundant facts – facts that are different from already extracted facts are added to the answer set

BBN Approach (cont.)

- Check if document contains target
  - First...Last for <u>who</u>, full match for <u>what</u>
  - Sentence match can be direct or through coreference; name match uses last name only
- Extract kernel facts
  - appositive and copula constructions
    - "George Bush, the president..." "Gearge Bush is the president..." (this is done using parsed sentences)

BBN Approach (cont.)

- Extract kernel facts
  - <u>special</u> and <u>ordinary</u> propositions: pred(role:arg,.....role:arg) for example love(subj:mary,obj:john) for "Mary loves John" – an special proposition would be "born in" of "educated in"
  - ~ 40 structured patterns typically used to define terms (TERM is NP)
  - Relations – 24 specific types of binary relations such as the staff of an organization
  - Full sentences used as fall back – do not match any of the above

BBN Approach (cont.)

- Ranking kernel facts
  - 1) appositives and copula ranked higher; 2) structured patterns; 3) special props; 4) relations; 5) props and sentences
  - Question profile: centroid of definitions from on-line dictionaries (e.g., Wikipedia); centroid of set of biographies; or centroid of all kernel facts
  - a similarity metric using tf*idf is used to rank the facts

BBN Approach (cont.)

- Redundancy removal
  - for propositions to be equivalent, same predicate and same argument head
  - for structured patterns, if the sentence was selected by a pattern used at least two times, then redundant
  - for other facts, check word overlap (>0.70 overlap is redundant)

BBN Approach (cont.)

- Algorithm for generating definitions
  - S={}
  - Rank all kernel facts based on profile similarity; iterate over the facts and discard redundant until there are m facts in S
  - Rank all remaining based on type (first) and similarity (second) add to S until maximum allowance reached or number of sentences and ordinary props greater than n
  - return S
- there is also a fall back approach when the above procedure does not produce any results – this is based on information retrieval

LREC 2008 – Marrakech - Morocco

Other Techniques

- Off-line strategies for identification in news paper articles of cases of <Concept, Instance> such as “Bush, President of the United States” (Fleishman&al’03)
  - use 2 types of patterns common noun (CN) proper noun (PN) constructions (English goalkeeper Seaman) and appositive constructions (Seaman, the English goalkeeper)
  - use a filter (classifier) to weed out noise
    - a number of features are used for the classifier including the pattern used; the semantic type of the head noun in the pattern; the morphology of the headnoun (e.g. spokes<u>man</u>); etc.

Other techniques

- DefScriber: definitional predicates and data-driven techniques (Blair-Goldensohn&al'03)
  - predicates = genus, species, non-specific – ML techniques over annotated corpus and patterns (manual)
  - centroid-based similarity and clustering

LREC 2008 – Marrakech - Morocco

Other techniques

- Best TREC QA 2006 def system used the Web to collect word frequencies (Kaisser'07)
  - Given a target obtain snippets from the web for queries containing the target words
  - Create a list of word frequencies
  - Retrieve docs from collection using target
  - Score sentences using the word frequencies
  - Pick up top ranked sentence and re-rank the rest of the sentences
  - Continue until termination

QA-definition approach (Saggion&Gaizauskas'04)

- linguistic patterns:
  - "*is a*" , "*such as*", "*consists of*", etc.
  - many forms in which definitions are expressed in texts
  - match definitions and non-definitions
  - "*Goth is a subculture*" & "*Becoming a Goth is a process that demands lots of effort*"

179

LREC 2008 – Marrakech - Morocco

QA-definition approach

- Secondary terms
  - Given multiple definitions of a specific definiendum, key defining terms are observed to recur across the definitions
  - For example
    - On the Web "*Goth*" seems to be associated with "*subculture*" in definition passages
  - Can we exploit known definitional contexts to assemble terms likely to co-occur with the definiendum in definitions?

180

Approach: use external sources

- Knowledge capture
  - identify definition passages (outside target collection) for the definiendum using patterns
  - WordNet, Wikipedia, Web in general
  - identify (secondary) terms associated to the definiendum in those passages
- During Answer extraction
  - use definiendum & secondary terms during IR
  - use secondary terms & patterns during IE from collection passages

Examples of Passages

Definiendum: aspirin

Pattern		Passage	
Uninstantiated	**Instantiated**	**Relevant**	**Not Relevant**
TERM is a	aspirin is a	Aspirin is a weak monotripic acid	Aspirin is a great choice for active people
such as TERM	such as aspirin	blood-thinners such as aspirin...	Look for travel size items such as aspirin
like TERM	like aspirin	non-steroidal antinflamatory drugs like aspirin	a clown is like aspirin, only he works twice as fast

Term List

- create a list of secondary terms
  - all WordNet terms, terms with count > 1 from web

Definiendum	WordNet	Encyclopedia	Web
aspirin	analgesic; anti-inflammatory; antipyretic; drug; …	inhibit; prostaglandin; ketofren; synthesis; …	drug; drugs; blood; ibuprofen; medication; pain; …
Aum Shirikyo	* NOTHING *	* NOTHING *	group; groups; cult; religious; japanese; etc.

LREC 2008 – Marrakech - Morocco

Definition extraction

- perform query expansion & retrieval
- analyse retrieved passages
  - look-up of definiendum, secondary terms, definition patterns
  - identify definition-bearing sentences
- identify answer
  - "*Who is Andrew Carnegie?*"
    - *In a question-and-answer session after the panel discussion, Clinton cited philanthropists from an earlier era such as Andrew Carnegie, J.P. Morgan, and John D. Rockefeller...*
    - *philanthropists from an earlier era such as Andrew Carnegie, J.P. Morgan, and John D. Rockefeller...*
- filter out redundant answers
  - vector space model and cosine similarity with threshold

Gold standard by NIST

Qid 1901: Who is Aaron Copland?

1901 1	vital	american composer
1901 2	vital	musical achievements ballets symphonies
1901 3	vital	born brooklyn ny 1900
1901 4	okay	son jewish immigrant
1901 5	okay	american communist
1901 6	okay	civil rights advocate
1901 7	okay	had senile dementia
1901 8	vital	established home for composers
1901 9	okay	won oscar for "the Heiress"
1901 10	okay	homosexual
1901 11	okay	teacher tanglewood music center boston symphony

185

LREC 2008 – Marrakech - Morocco

Evaluation

- NIST
  - matching system answers to human answers
- Metrics
  - « nugget recall » (NR) ~ traditional recall
  - « nugget precision » (NP) ~ space used by system answer is important
    - it is better to save space
  - « F-score » (F) harmonic mean of NR and NP where NR is 5 times more important than NP

186

Metrics

answer set is the set of nuggets your system returned

gold standard set if the set of nuggets identified by the assessors

$$NR = \frac{\#\text{ of essential nuggets returned}}{\#\text{ of essential nuggets}}$$

$allowance = 100 * (\#$ of essential and non essential nuggets returned)

$length =$ number of non - white - space characters in answer set

$NP = 1$ if $length < allowance$

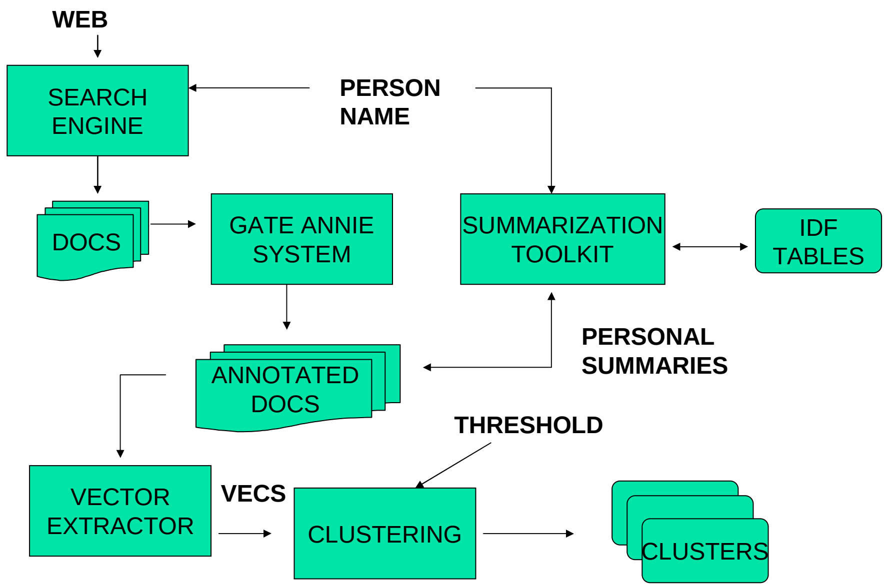
$$NP = 1 - \frac{length - allowance}{lenght} \text{ if } length \geq allowance$$

$$F = \frac{(1+\alpha^2) * NP * NR}{\alpha^2 * NP * NR}$$

LREC 2008 – Marrakech - Morocco

Creation of person profiles

- Creation of person profiles assume that the input set of documents refer to a unique individual (Who is X?)
- Summaries can be used to cluster documents referring to a single individual and each cluster can be summarized in its own right
  - X the scientist; X the politician; X the artist; etc.



Clustering

Given a set of documents and a threshold

1. Initially there are as many clusters as documents
2. All clusters are compared using a similarity metric
3. At each iteration the two most similar clusters are merged if their similarity is greater than a threshold (otherwise stop and return clusters)
4. Continue with step 2

Document Representation

- term frequency (tf) of term t in document d = the number of times t occurs in d
- inverted document frequency (idf) of term t in collection c = the number of documents in c containing t
- Bag-of-word approach = words are terms
  - text = ($word_1$=w_1....)
- Semantic-based approach = named entities are terms (person, location, organization, date, address)
  - text = (ne_1=w_1....)
- Extract terms from document summaries

Document Representation

- local IDF tables are computed for each set of documents
- weights are tf*log(N/idf) – N is the size of the document set
- simC is the cluster similarity; simD is the document similarity which is the cosine metric

$$sim_C(C_1, C_2) = \max_{d_i \in C_1;\, d_j \in C_2} sim_D(d_i, d_j)$$

- threshold estimated over training data
  - the algorithm is run over the training and the similarity value for the optimal f-score noted for each instance
  - the threshold is taken as the average of the optimal thresholds

Possible Summaries

- Coreference chains associated to the target person name are identified (in each document)
- All elements in a coreference chain containing the target person are marked
- Sentences containing marked person name are selected for summary
  - On Tuesday, Hobbs was arrested on murder charges in the Mother's Day stabbings of his 8-year-old daughter and the little girl's best friend, who were killed after they went biking in a park.
  - Jerry Hobbs said he resigned from the Temecula Valley school board, in part, because other trustees would not consider switching from trimesters to semesters in high schools.

Possible Summaries

- Sentences containing biographical patterns involving the target person name are selected for a summary
  - Patterns
    - target (is|…) (a|…) dp
    - target's….
    - target, who…
  - Sentences with patterns
    - Jerry Hobbs, who was recently released….
    - Hobbs, 34, was questioned through…
    - Jerry Hobbs is a research professor at ….

Background Gathering Application – Summarization and QA (Gaizauskas&al07)

- *Background gathering*: the task of collecting information from the news wire and other archives to contextualise and support a breaking news story
- Backgrounder components
  - similar events in the past; role players' profiles; factual information on the event
- Collaboration with Press Association
  - 11 year archive with more than 8 million stories
- Information access system comprising: Information Retrieval, Text Summarization, Question Answering, "Similar Event Search"

195

Background Examples

- Breaking News

  "Powerful earthquake shook Turkey today"
- Past Similar Events

  "Last year an earthquake measuring 6.3-magnitude hit southern Turkey killing 144 people."
- Extremes

  "Europe's biggest quake hit Lisbon, Portugal, on November 1, 1755, when 60,000 people died as the city was devastated and giant waves 10 metres high swept through the harbour and on to the shore."
- Definitions

  "Quakes occur when the Earth's crust fractures, a process that can be caused by volcanic activity, landslides or subterranean collapse. The resulting plates grind together causing the tremors.

196

Text Analysis Resources

- General Architecture for Text Engineering
  - (http://gate.ac.uk)
  - Tokenisation, Sentence Identification, POS tagging, NE recognition, etc.
- SUPPLE syntactic-semantic Parser
  - (http://nlp.shef.ac.uk/research/supple)
  - syntactic parsing and creation of logical forms
- Summarization Toolkit
  - (http://www.dcs.shef.ac.uk/~saggion)
  - Single and multi document summarization
- Lucene
  - (http://lucene.apache.org)
  - Text indexing and retrieval

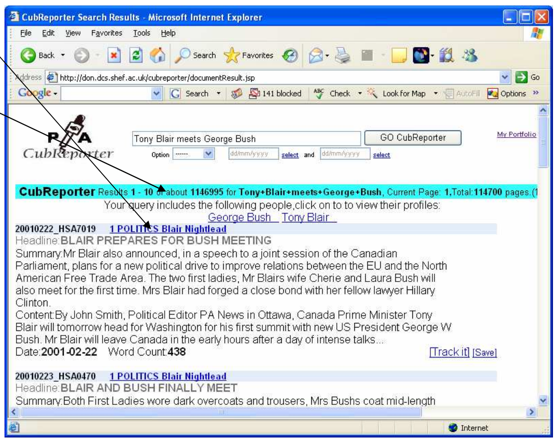
Finding Stories

auto summaries

profiles

metadata

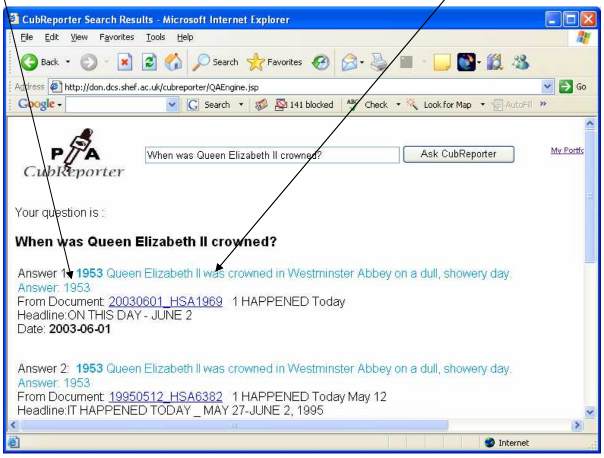
stories



Getting Answers

answers

context

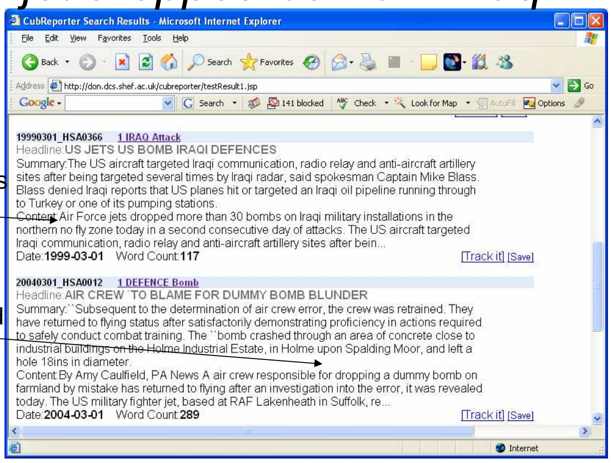


Getting Similar Events

“jet dropped bomb in Iraq”

jets drop bombs

bombs dropped



Some Research Topics

- Multi-sentence non-extractive summarization – beyond headline generation
- "State-of-art" summaries – what is the state of the art on topic x?
- "Background" summaries for a given story
- Adaptable summarization – create a system to summarize event X and techniques to adapt the system to event Y
- Summarize opinions about topic X (person, event, etc.)

APPENDIX

Abstractor's at work (Endres-Niggemeyer'95)

- systematic study of professional abstractors
- "speak-out-loud" protocols
- discovered operations during document condensation
  - use of document structure
  - top-down strategy + superficial features
  - cut-and-paste

Abstract's structure (Liddy'91)

- Identification of a text schema (grammar) of abstracts of empirical research
- Identification of lexical clues for predicting the structure
- From abstractors to a linguistic model
  - ERIC and PsycINFO abstractors as subjects of experimentation

Abstract's structure

- Three levels of information
  - proto-typical
    - hypothesis; subjects; conclusions; methods; references; objectives; results
  - typical
    - relation with other works; research topic; procedures; data collection; etc.
  - elaborated-structure
    - context; independent variable; dependent variable; materials; etc.
- Suggests that types of information can be identified based on "cue" words/expressions
- Many practical implications for IR systems

Finding source sentences (Saggion&Lapalme'02)

Source document	Abstract
In this paper we have presented a more efficient distributed algorithm which constructs a breadth-first search tree in an asynchronous communication network.	Presents a more efficient distributed breadth-first search algorithm for an asynchronous communication network.
We present a model and give an overview of related research.	Presents a model and gives an overview of related research.
We analyse the complexity of our algorithm and give some examples of performance on typical networks.	Analyses the complexity of the algorithm and gives some examples of performance on typical networks.

Document structure for abstracting

Title	2%
Author abstract	15%
First section	34%
Last section	3%
Headings and captions	33%
Other sections	13%

Keyword method: Luhn'58

- words which are frequent in a document indicate the topic discussed
- stemming algorithm ("systems" = "system")
- ignore "stop words" (i.e."the", "a", "for", "is")
- compute the distribution of each word in the document (tf)

Keyword method

- compute distribution of words in corpus (i.e., collection of texts)
- inverted document frequency

$$idf(term) = \log\left(\frac{NUMDOC}{NUMDOC(term)}\right)$$

$NUMDOC$ #docs in corpus

$NUMDOC(term)$ #docs where term occurs

209

LREC 2008 – Marrakech - Morocco

Keyword method

- consider only those terms such that tf*idf > thr
- identify clusters of keywords
  - $[X_i \; X_{i+1} \ldots X_{i+n-1}]$
- compute weight
- normalize

$$\frac{\#significant(C)^2}{\#words(C)}$$

$$weight(t) = tf(t).ifd(t)$$

$$weight(S) = \sum_{t \in S} weight(t)$$

210

Position: Edmundson'69

- Important sentences occur in specific positions
  - "lead-based" summary (Brandow'95)
  - inverse of position in document works well for the "news"

$$position(S_i) = (i)^{-1}$$

- Important information occurs in specific sections of the document (introduction/conclusion)

Position

- Extra points for sentences in specific sections
  - make a list of important sections

    LIST= "introduction", "method", "conclusion", "results", ...

- Position evidence (Baxendale'58)
  - first/last sentences in a paragraph are topical
  - give extra points to = initial | middle | final

Position

- Position depends on type of text!
- "Optimum Position Policy" (Lin & Hovy'97) method to learn "positions" which contain relevant information OPP= { (p1,s2), (p2,s1), (p1,s1), ...}
  - pi = paragraph num; si = sentence num
  - "learning" method uses documents + abstracts + keywords provided by authors
  - average number of keywords in the sentence
  - 30% topic not mentioned in text
  - title contains 50% topics
  - title + 2 best positions 60% topics

LREC 2008 – Marrakech - Morocco

Title method: Edmundson'69

- Hypothesis: title of document indicates its content

TITLE:
IBM's statistical question answering system - TREC-11
SENTENCE:
In this paper, we document our efforts to extend our statistical question answering system for TREC-11.

- therefore, words in title help find relevant content
- create a list of title words, remove "stop words"

$$title(S) = |TIT \cap S|$$

LREC 2008 – Marrakech - Morocco

Cue method: Edmundson'69;Paice'81

- Important sentences contain cue words/indicative phrases
  - "The main aim of the present paper is to describe…" (IND)
  - "The purpose of this article is to review…" (IND)
  - "In this report, we outline…" (IND)
  - "Our investigation has shown that…" (INF)
- Some words are considered bonus others stigma
  - bonus: comparatives, superlatives, conclusive expressions, etc.
  - stigma: negatives, pronouns, etc.

215

LREC 2008 – Marrakech - Morocco

FRUMP

- Knowledge structure = sketchy-scripts, adaptation of Shank & Abelson scripts (1977)
- sketchy-scripts contain only the relevant information of an event
- ~50 sketchy-scripts manually developed for FRUMP
- Interpretation is based on skimming

216

FRUMP

- When a key word is found one or more scripts are activated
- The activated scripts guide text interpretation, syntactic analysis is called on demand
- When more than one script is activated, heuristics decide which represents the correct interpretation
- Because the representation is language-independent, it can be used to generate summaries in various languages

LREC 2008 – Marrakech - Morocco

FRUMP

- Evaluation: one day of processing text
- 368 stories
  - 100 not news articles
  - 147 not of the script type
  - 121 could be understood
  - for 29 FRUMP has scripts
  - only 11 were processed correctly + 2 almost correctly = 3% correct; on average 10% correct
- problems
  - incorrect variable binding
  - could not identify script
  - incorrect script used to interpret (no script)
  - incorrect script used to interpret (correct script present)

FRUMP

- 50 scripts is probably not enough for interpreting most stories
- knowledge was manually coded
- how to learn new scripts

Vatican City. The dead of the Pope shakes the world. He passed away…

Earthquake in the Vatican. One dead.

LREC 2008 – Marrakech - Morocco

Extracts by lexical chains

- Compute the contribution of N to C as follows
  - If C is empty consider the relation to be “repetition” (identity)
  - If not identify the last element M of the chain to which N is related
  - Compute distance between N and M in number of sentences (1 if N is the first word of chain)
  - Contribution of N is looked up in a table with entries given by type of relation and distance
    - e.g., hyper & distance=3 then contribution=0.5

Extracts by lexical chains

- After inserting all nouns in chains there is a second step
- For each noun, identify the chain where it most contributes; delete it from the other chains and adjust weights
- Select sentences that belong or are covered by "strong chains"

Extracts by lexical chains

- Strong chain:
  - weight(C) > thr
  - thr = average(weight(Cs)) + 2*sd(weight(Cs))
- selection:
  - H1: select the first sentence that contains a member of a strong chain
  - H2: select the first sentence that contains a "representative" member of the chain
  - H3: identify a text segment where the chain is highly dense (density is the proportion of words in the segment that belong to the chain)

Headline generation

- Content selection
  - What document features influence the words of the headline
  - A possible feature: the words of the document
    - W is in summary & W is in document
  - This feature can be computed as

$$p(w_i \in T | w_i \in D) = \frac{p(w_i \in D | w_i \in T).p(w_i \in T)}{p(w_i \in D)}$$

  - Other feature: how many words to select?

$$p(len(T) = n)$$

  - Easiest solution is to use a fixed length per document type

Headline generation

- Surface realization
  - Compute the probability of observing $w_1 \ldots w_n$

$$\prod p(w_i | w_1 \ldots . w_{i-1})$$

  - 2-grams approximation

$$\prod p(w_i | w_{i-1})$$

Headline generation

- Model combination
  - we want the best sequence of words

$$p(w_1 \ldots w_n) = \begin{array}{l} \prod p(w_i \in T \mid w_i \in D) * \\ p(len(T) = n) * \\ \prod p(w_i \mid w_1 \ldots . w_{i-1}) \end{array}$$

content model

realization model

LREC 2008 – Marrakech - Morocco

Headline generation

- Search using the following formula (note the use logarithm)

$$\operatorname{argmax}_T(\alpha \sum \log(p(w_i \in T | w_i \in D)) +$$
$$\beta \log(p(lon(T) = n)) +$$
$$\gamma \sum \log(w_i | w_{i-1}))$$

- can be used to find the best sequence
- One has to consider the problem of data sparseness
  - words never seen
  - 2-grams never seen
- There are "smoothing" and "back-off" models to deal with the problems

Headline Generation: Evaluation

- Compare automatic headline with original headline
  - Words in common
- Various lengths evaluated
  - 4 words give acceptable results (?) 1 out of 5 headlines contain all words of the original
- Grammaticality is an issue, however headlines have their own syntax
- Other features
  - POS & position

Cut & Paste human examples

Example 1: add description for people or organization

Original Sentences:

Sentence 34: "We're trying to prove that there are big benefits to the patients by involving them more deeply in their treatment", said Paul Clayton, chairman of the dept. dealing with computerized medical information at Columbia.

Sentence 77: "The economic payoff from breaking into health care records is a lot less than for banks", said Clayton at Columbia.

Rewritten Sentences:

Combined: "The economic payoff from breaking into health care records is a lot less than for banks", said Paul Clayton, chairman of the dept. dealing with computerized medical information at Columbia.

Example 2: extract common elements

Original Sentences:

Sentence 8: but it also raises serious questions about the privacy of such highly personal information wafting about the digital world

Sentence 10: The 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

Rewritten Sentences :

Combined: but it also raises the issue of privacy of such personal information and this issue hits the head on the nail in the broader debate about privacy and security on the internet.

Cut&Paste human examples

Example 3: reduce and join sentences by adding connectives or punctuations

Original Sentences:

Sentence 7: Officials said they doubted that Congressional approval would be needed for the changes, and they forsaw no barriers at the Federal level.

Sentence 8: States have wide control over the availability of methadone, however.

Rewritten Sentences :

Combined: Officials said they foresaw no barriers at the Federal level; however, States have wide control over the availability of methadone.

Example 4: reduce and change one sentence to a clause

Original Sentences:

Sentence 25: in GPI, you specify an RGB COLOR value with a 32-bit integer encoded as follows: 00000000* Red * Green * Blue The high 8 bits are set to 0.

Sentence 27: this encoding scheme can represent some 16 million colors

Rewritten Sentences :

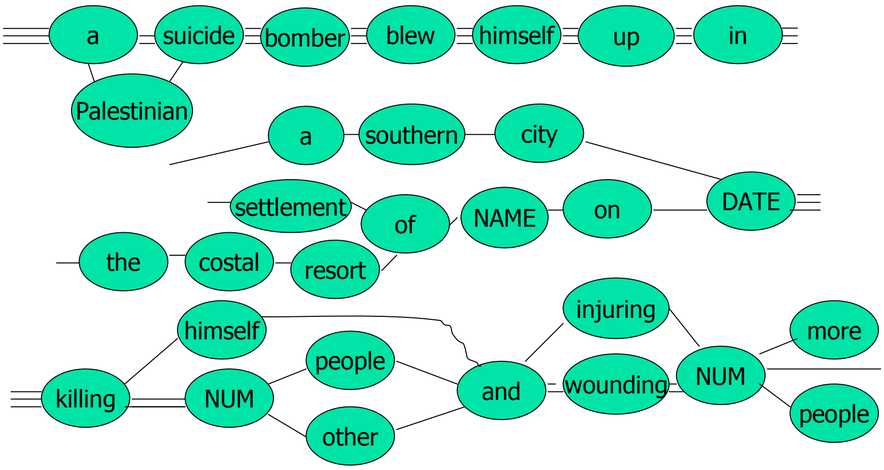
Combined: GPI describes RGB colors as 32-bit integers that can describe 16 million colors

LREC 2008 – Marrakech - Morocco

Paraphrase

- apply a multi-sequence alignment algorithm to represent paraphrases as lattices
- identify arguments (variable) as zones of great variability in the lattices
- generation of paraphrases can be done by matching against the lattices and generating as many paraphrases as paths in the lattice

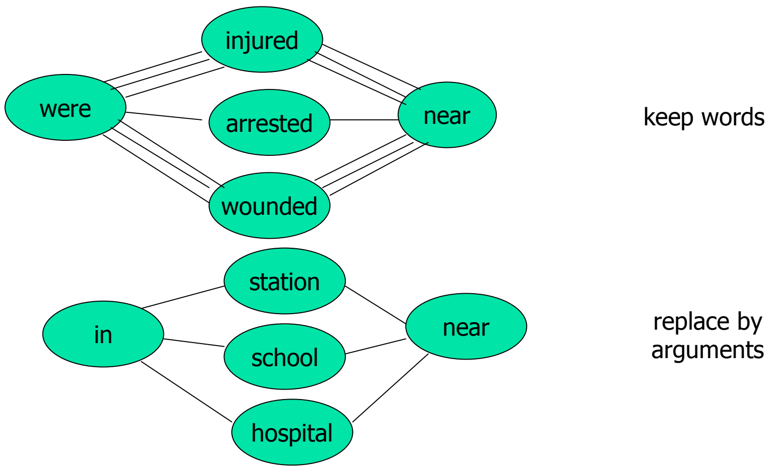
Lattices and backbones



231

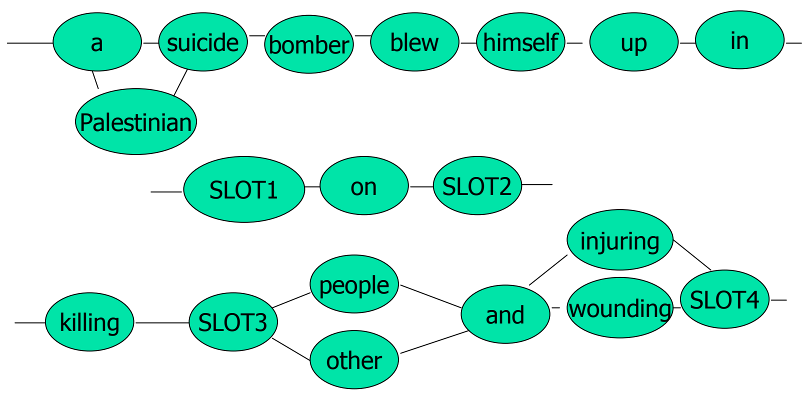
LREC 2008 – Marrakech - Morocco

Arguments or Synonyms?



232

Patterns induced



Generating paraphrases

- finding equivalent patterns
  - X injured Y people, Z seriously = Y were injured by X among them Z were in serious condition
- exploit the corpus
  - equivalent patterns will have similar arguments/slots in the corpus
  - given two clusters from where the patterns were derived identify sentences "published" on the same date & topic
  - compare the arguments in the pattern variables
  - patterns are equivalent if overlap of word in arguments > thr

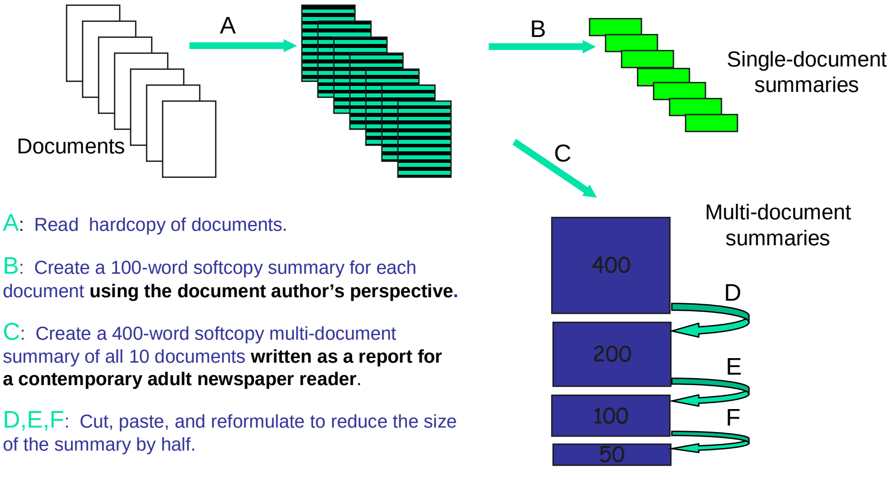
DUC 2001

- Task 1
  - given a document, create a generic summary of the document (100 words)
  - 30 sets of ~10 documents each
- Task 2
  - given a set of documents, create summaries of the set (400, 200, 100, 50 words)
  - 30 sets of ~ 10 documents each

235

SLIDE FROM Document Understanding Conferences

Human summary creation



A: Read hardcopy of documents.

B: Create a 100-word softcopy summary for each document **using the document author's perspective.**

C: Create a 400-word softcopy multi-document summary of all 10 documents **written as a report for a contemporary adult newspaper reader**.

D,E,F: Cut, paste, and reformulate to reduce the size of the summary by half.

236

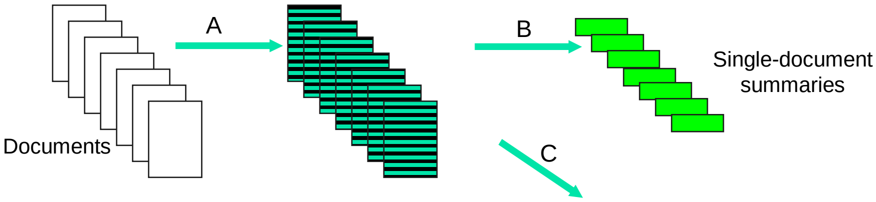
DUC 2002

- Task 1
  - given a document, create a generic summary of the document (100 words)
  - 60 sets of ~10 documents each
- Task 2
  - given a set of documents, create summaries of the set (400, 200, 100, 50 words)
  - given a set of documents, create two extracts (400, 200 words)
  - 60 sets of ~ 10 documents each

LREC 2008 – Marrakech - Morocco

SLIDE FROM Document Understanding Conferences

Human summary creation

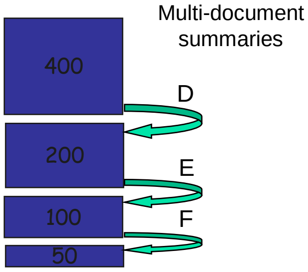


A: Read hardcopy of documents.

B: Create a 100-word softcopy summary for each document **using the document author's perspective.**

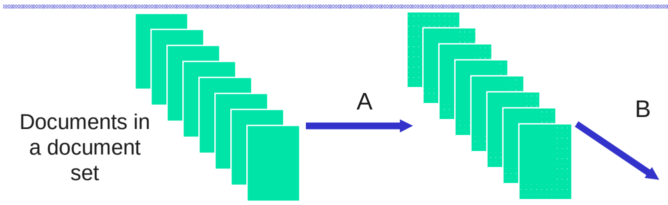
C: Create a 400-word softcopy multi-document summary of all 10 documents **written as a report for a contemporary adult newspaper reader**.

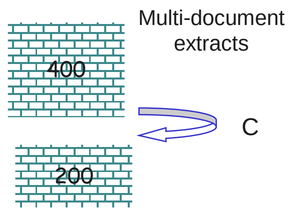
D,E,F: Cut, paste, and reformulate to reduce the size of the summary by half.



SLIDE FROM Document Understanding Conferences

Manual extract creation





A: **Automatically tag sentences**

B: **Create a 400-word softcopy multi-document extract of all 10 documents together**

C: **Cut and paste to produce a 200-word extract**

239

DUC 2003

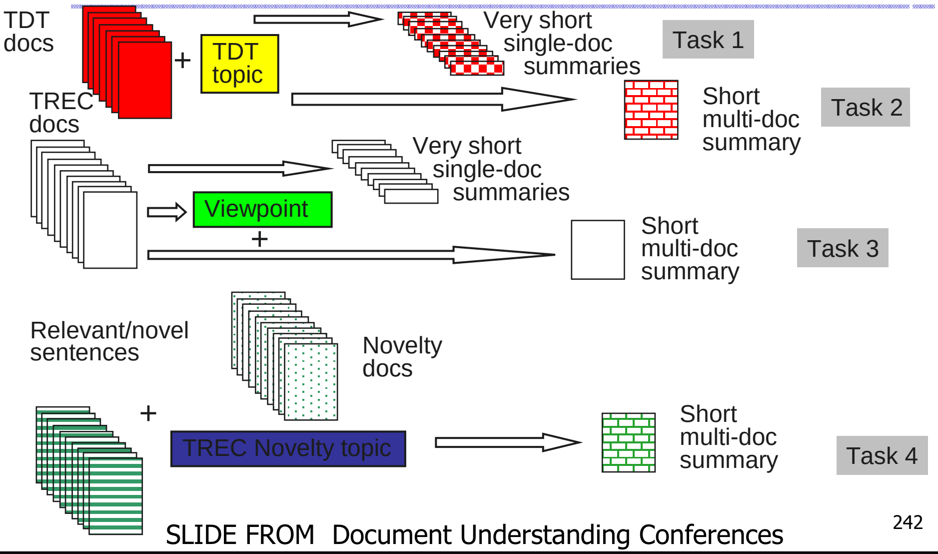
- Task 1
  - 10 words single-document summary
- Task 2
  - 100 word multi-document summary of cluster related by an event
- Task 3
  - given a cluster and a viewpoint, 100 word multi-document summary of cluster
- Task 4
  - given a cluster and a question, 100 word multi-document summary of cluster

240

Viewpoints & Topics & Questions

Viewpoint:
Forty years after poor parenting was thought to be the cause of schizophrenia, researchers are working in many diverse areas to refine the causes and treatments of this disease and enable early diagnosis.
Topic:
30042 - PanAm Lockerbie Bombing Trial
Seminal Event
WHAT: Kofi Annan visits Libya to appeal for surrender of PanAm bombing suspects
WHERE: Tripoli, Libya
WHO: U.N. Secretary-General Kofi Annan; Libyan leader Moammar Gadhafi
WHEN: December, 1998
Question:
What are the advantages of growing plants in water or some substance other than soil?

Manual abstract creation



SLIDE FROM Document Understanding Conferences

Single-document summary (DUC)

<SUM DOCSET="d04" TYPE="PERDOC" SIZE="100" DOCREF="FT923 6455" SELECTOR="A" SUMMARIZER="A">
US cities along the Gulf of Mexico from Alabama to eastern Texas were on alert last night as Hurricane Andrew headed west after hitting southern Florida leaving at least eight dead, causing severe property damage, and leaving 1.2 million homes without electricity. Gusts of up to 165 mph were recorded. It is the fiercest hurricane to hit the US in decades. As Andrew moved across the Gulf there was concern that it might hit New Orleans, which would be particularly susceptible to flooding, or smash into the concentrated offshore oil facilities. President Bush authorized federal disaster assistance for the affected areas.</SUM>

243

LREC 2008 – Marrakech - Morocco

Multi-document summaries (DUC)

<SUM DOCSET="d04" TYPE="MULTI" SIZE="50" DOCREF="FT923-5267 FT923-6110 FT923-6455 FT923-5835 FT923-5089 FT923-5797 FT923-6038" SELECTOR="A" SUMMARIZER="A">
Damage in South Florida from Hurricane Andrew in August 1992 cost the insurance industry about $8 billion making it the most costly disaster in the US up to that time. There were fifteen deaths and in Dade County alone 250,000 were left homeless.</SUM>

<SUM DOCSET="d04" TYPE="MULTI" SIZE="100" DOCREF="FT923-5267 FT923-6110 FT923-6455 FT923-5835 FT923-5089 FT923-5089 FT923-5797 FT923-6038" SELECTOR="A" SUMMARIZER="A">
Hurricane Andrew which hit the Florida coast south of Miami in late August 1992 was at the time the most expensive disaster in US history. Andrew's damage in Florida cost the insurance industry about $8 billion. There were fifteen deaths, severe property damage, 1.2 million homes were left without electricity, and in Dade county alone 250,000 were left homeless. Early efforts at relief were marked by wrangling between state and federal officials and frustrating delays, but the White House soon stepped in, dispatching troops to the area and committing the federal government to rebuilding and funding an effective relief effort.</SUM>

244

Extracts (DUC)

<SUM DOCSET="d061" TYPE="MULTI-E" SIZE="200"
DOCREF="AP880911-0016 AP880912-0137 AP880912-0095 AP880915-0003 AP880916-0060
WSJ880912-0064" SELECTOR="J" SUMMARIZER="B">
<s docid="WSJ880912-0064" num="18" wdcount="15"> Tropical Storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night.</s>
<s docid="AP880912-0137" num="22" wdcount="13"> Gilbert reached Jamaica after skirting southern Puerto Rico, Haiti and the Dominican Republic.</s>
<s docid="AP880915-0003" num="13" wdcount="33"> Hurricane Gilbert, one of the strongest storms ever, slammed into the Yucatan Peninsula Wednesday and leveled thatched homes, tore off roofs, uprooted trees and cut off the Caribbean resorts of Cancun and Cozumel.</s>
<s docid="AP880915-0003" num="44" wdcount="21"> The Mexican National Weather Service reported winds gusting as high as 218 mph earlier Wednesday with sustained winds of 179 mph.</s>

DUC 2004 – Some systems

- Task 1
  - TOPIARY (Zajic&al'04)
    - University of Maryland; BBN
    - Sentence compression from parse tree
    - Unsupervised Topic Discovery (UTD): statistical technique to associate meaningful names to topics
    - Combination of both techniques
  - MEAD (Erkan&Radev'04)
    - University of Michigan
    - Centroid + Position + Length
    - Select one sentence as S sumary

DUC 2004 – Some systems

- Task 2
  - CLASSY (Conroy&al'04)
    - IDA/Center for Computing Sciences; Department of Defence; University of Maryland
    - HMM with summary and non-summary states
      - Observation input = topic signatures
    - Co-reference resolution
    - Sentence simplification
  - Cluster Relevance & Redundancy Removal (Saggion&Gaizauskas'04)
    - University of Sheffield
    - Sentence cluster similarity + sentence lead document similarity + absolute position
    - N-gram based redundancy detection

247

DUC 2004 – Some systems

- Task 3
  - LAKHAS (Douzidia&Lapalme'04)
  - Universite de Montreal
  - Summarize from Arabic documents, then translates
  - Sentence scoring= lead + title + cue + tf*idf
  - Sentence reduction = name substitution; word removal; phrase removal; etc.
  - After translation with Ajeeb (commercial system) good results
  - After translation with ISI best system

248

DUC 2004 – Some systems

- Task 5
  - Lite-GISTexter (Lacatusu&al'04)
  - Language Computer Corporation
  - Syntactic structure
    - entity in appositive construction ("X, a …")
    - entity subject of copula ("X is the…")
    - sentence containing key are scored by syntactic features

References

[1] Proceedings of the SIGIR Workshop on Information Retrieval for Question Answering (IR4QA). Sheffield, UK, 2004. Available, September 2004, from `http://nlp.shef.ac.uk/ir4qa04/`.

[2] AAAI. *Intelligent Text Summarization Symposium, AAAI 1998 Spring Symposium Series. March 23-25*, Stanford, USA, March 23-25 1998.

[3] Steven Abney, Michael Collins, and Amit Singhal. Answer Extraction. In *Proceedings of ANLP 2000*, 2000.

[4] ACL. *Workshop on Automatic Summarization, ANLP-NAACL2000, April 30*, Seattle, Washington, USA, April 30 2000.

[5] ACL/EACL. *Workshop on Intelligent Scalable Text Summarization*, ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization,11 July 1997, Madrid, Spain, 1997.

[6] AFNOR. *Recommandations aux Auteurs des Articles Scientifiques et Techniques pour la Rédaction des Résumés.* Association Française de Normalisation, 1984.

[7] Eugene Agichtein, Steve Lawrence, and Luis Gravano. Learning Search Engine Specific Query Transformations for Question Answering. In *Proceedings of the 10th International World Wide Web Conference (WWW10)*, Hong Kong, May 1-5 2001.

[8] Richard Alterman. Text Summarization. In S.C. Shapiro, editor, *Encyclopedia of Artificial Intelligence*, volume 2, pages 1579–1587. Jonh Wiley & Sons, Inc., 1992.

[9] Richard Alterman and Lawrence A. Bookman. Some Computational Experiments in Summarization. *Discourse Processes*, 13:143–174, 1990.

[10] ANSI. *Writing Abstracts.* American National Standards Institute, 1979.

[11] D.E. Appelt, J.R. Hobbs, J. Bear, D. Israel, and M. Tyson. Fastus: A finite-state processor for information extraction from real-world text. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI'93)*, volume 2, pages 1172–1178, 1993.

[12] R. Barzilay, N. Elhadad, and K. McKeown. Sentence ordering in multi-document summarization. 2001.

[13] Regina Barzilay and Michael Elhadad. Using Lexical Chains for Text Summarization. In *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 10–17, Madrid, Spain, July 1997.

[14] Regina Barzilay, Kathleen McKeown, and Michael Elhadad. Information Fusion in the Context of Multi-Document Summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 550–557, Maryland, USA, 20-26 June 1999.

[15] Barzilay, R. and Lee, L. Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization. In *Proceedings of HLT-NAACL 2004*, 2004.

[16] P.B. Baxendale. Man-made Index for Technical Litterature - an experiment. *IBM J. Res. Dev.*, 2(4):354–361, 1958.

[17] M. Benbrahim and K. Ahmad. Text Summarisation: the Role of Lexical Cohesion Analysis. *The New Review of Document & Text Management*, pages 321–335, 1995.

[18] Charles L. Bernier. Abstracts and Abstracting. In E.D. Dym, editor, *Subject and Information Analysis*, volume 47 of *Books in Library and Information Science*, pages 423–444. Marcel Dekker, Inc., 1985.

[19] J. Berri. Mise en œuvre de la Méthode d'Exploration Contextuelle pour le Résumé Automatique de Textes. Implémentation du Système SERAPHIN. In P. Bouffard and A. Kharrat, editors, *Actes du Premier Colloque Étudiant de Linguistique Informatique de Montréal (CLIM 96)*, pages 128–135, 1996.

[20] Matthew W. Bilotti, Boris Katz, and Jimmy Lin. What Works Better for Question Answering: Stemming or Morphological Query Expansion. In *Proceedings of the SIGIR Workshop on Information Retrieval for Question Answering (IR4QA)*, Sheffield, UK, July 29 2004.

[21] E.J. Black and Frances C. Johnson. A practical evaluation of two rule-based automatic abstracting techniques. *Expert System for Information Management*, 1(3):159–177, 1988.

[22] William J. Black. Knowledge based abstracting. *Online Review*, 14(5):327–340, 1990.

[23] Sasha Blair-Goldensohn, Kathleen R. McKeown, and Andrew Hazen Schlaikjer. DefScriber: A Hybrid Approach for Answering Definitional Questions (Demo). In *Processdings of the 26th ACM SIGIR Conference*, Toronto, Canada, July 2003. ACM.

[24] H. Borko and C. Bernier. *Abstracting Concepts and Methods*. Academic Press, 1975.

[25] Ronald Brandow, K. Mitze, and Lisa F. Rau. Automatic Condensation of Electronic Publications by Sentence Selection. *Information Processing & Management*, 31(5):675–685, 1995.

[26] Eric Brill, Jimmy Lin, Michele Banko, Susan Dumais, and Andrew Ng. Data-Intensive Question Answering. In *Proceedings of the Tenth Text REtrieval Conference*, 2001.

[27] John D. Burger, Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, Sanda Harabagiu, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George A. Miller, Dan Moldovan, Bill Ogden, John Prager, Ellen Riloff, Amit Singhal, Rohini Shrihari, Tomek Strzalkowski, Ellen M. Voorhees, and Ralph Weischedel. Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A). Technical report, NIST, 2000. Available, September 2002, from `http://www-nlpir.nist.gov/projects/duc/roadmapping.html`.

[28] Jaime G. Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Research and Development in Information Retrieval*, pages 335–336, 1998.

[29] E. Cartier. Analyse automatique des textes: l'exemple des informations définitoires. In *Rencontre Internationale sur l'extraction le Filtrage et le Résumé Automatique*, pages 6–18, Novembre 1998.

[30] D.B. Cleveland and A.D. Cleveland. *Introduction to Indexing and Abstracting*. Libraries Unlimited, Inc., 1983.

[31] Paul Clough, Robert Gaizauskas, Scott Piao, and Yorick Wilks. METER: MEasuring TExt Reuse. In *Proceedings of the ACL*. Association for Computational Linguistics, July 2002. To appear.

[32] Eduard T. Cremmins. *The Art of Abstracting*. ISI PRESS, 1982.

[33] Hang Cui, Min-Yen Kan, and Tat-Seng Chua. Unsupervised Learning of Soft Patterns for Definitional Question Answering. In *Proceedings of the Thirteenth World Wide Web Conference (WWW 2004)*, New York, May 2004.

[34] Hang Cui, Min-Yen Kan, Tat-Seng Chua, and Jing Xiao. A Comparative Study on Sentence Retrieval for Definitional Question Answering. In *Proceedings of the SIGIR Workshop on Information Retrieval for Question Answering (IR4QA)*, Sheffield, UK, July 29 2004.

[35] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002.

[36] H. Dalianis, M. Hassel, K. de Smedt, A. Liseth, T.C. Lech, and J. Wedekind. Porting and evaluation of automatic summarization. In *Nordisk Sprogteknologi*, pages 107–121, 2004.

[37] Gerald DeJong. An Overview of the FRUMP System. In W.G. Lehnert and M.H. Ringle, editors, *Strategies for Natural Language Processing*, pages 149–176. Lawrence Erlbaum Associates, Publishers, 1982.

[38] R.L. Donaway, K.W. Drummey, and L.A. Mather. A Comparison of Rankings Produced by Summarization Evaluation Measures. In *Proceedings of the Workshop on Automatic Summarization, ANLP-NAACL2000*, pages 69–78. Association for Computational Linguistics, 30 April 2000 2000.

[39] H.P. Edmundson. New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery*, 16(2):264–285, April 1969.

[40] B. Endres-Niggemeyer and K. Sparck-Jones, editors. *Information Processing & Management Special Issue on Text Summarization*, volume 31. Pergamon, 1995.

[41] Brigitte Endres-Niggemeyer. SimSum: an empirically founded simulation of summarizing. *Information Processing & Management*, 36:659–682, 2000.

[42] Brigitte Endres-Niggemeyer, E. Maier, and A. Sigel. How to Implement a Naturalistic Model of Abstracting: Four Core Working Steps of an Expert Abstractor. *Information Processing & Management*, 31(5):631–674, 1995.

[43] Brigitte Endres-Niggemeyer, W. Waumans, and H. Yamashita. Modelling Summary Writting by Introspection: A Small-Scale Demonstrative Study. *Text*, 11(4):523–552, 1991.

[44] D.K. Evans, J.L. Klavans, and K.R. McKeown. Columbia Newsblaster: Multilingual News Summarization on the Web. In *Proceedings of NAACL/HLT*, 2004.

[45] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database.* The MIT Press, 1998.

[46] Thérèse Firmin and Michael J. Chrzanowski. An Evaluation of Automatic Text Summarization Systems. In I. Mani and M.T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 325–336. The MIT Press, 1999.

[47] Michael Fleischman, Eduard Hovy, and Abdessamad Echihabi. Offline Strategies for Online Question Answering: Answering Questions Before They Are Asked. In *Proceedings of the ACL 2003*, pages 1–7. ACL, 2003.

[48] R. Gaizauskas, M. Hepple, H. Saggion, and M. Greenwood. SUPPLE: A Practical Parser for Natural Language Engineering Applications. In *International Workshop on Parsing Technologies*, 2005.

[49] R. Gaizauskas, H. Saggion, and E. Barker. Information Access and Natural Language Processing: A Stimulating Dialogue. In K. Ahmad, C. Brewster, and M. Stevenson, editors, *Words and Intelligence II: Essays in Honor of Yorick Wilks*. Springer, 2007.

[50] Robert Gaizauskas, Mark A. Greenwood, Mark Hepple, Ian Roberts, and Horacio Saggion. The University of Sheffield's TREC 2004 Q&A Experiments. In *Proceedings of the 13th Text REtrieval Conference*, 2004.

[51] Robert Gaizauskas, Mark A. Greenwood, Mark Hepple, Ian Roberts, Horacio Saggion, and Matthew Sargaison. The University of Sheffield's TREC 2003 Q&A Experiments. In *Proceedings of the 12th Text REtrieval Conference*, 2003.

[52] Timothy R. Gibson. *Towards a Discourse Theory of Abstracts and Abstracting*. Department of English Studies. University of Nottingham, 1993.

[53] Jade Goldstein, Vibhu O. Mittal, Jaime G. Carbonell, and Mark Kantrowitz. Multi-document summarization by sentence extraction. In *Proceedings of ANLP/NAACL workshop on Automatic Summmarization*, Seattle, WA, April 2000.

[54] Pamela Grant. *The Integration of Theory and Practice in the Development of Summary-Writting Strategies*. PhD thesis, Université de Montréal. Faculté des études supérieures., 1992.

[55] Bert F. Green, Alice K. Wolf, Carol Chomsky, and Kenneth Laughery. BASEBALL: An Automatic Question Answerer. In *Proceedings of the Western Joint Computer Conference 19*, pages 219–224, 1961.

[56] Mark A. Greenwood. AnswerFinder: Question Answering from your Desktop. In *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK '04)*, pages 75–80, Birmingham, UK, January 7 2004.

[57] Mark A. Greenwood and Robert Gaizauskas. Using a Named Entity Tagger to Generalise Surface Matching Text Patterns for Question Answering. In *Proceedings of the Workshop on Natural Language Processing for Question Answering (EACL03)*, pages 29–34, Budapest, Hungary, April 14 2003.

[58] Mark A. Greenwood, Ian Roberts, and Robert Gaizauskas. The University of Sheffield TREC 2002 Q&A System. In *Proceedings of the 11th Text REtrieval Conference*, 2002.

[59] Mark A. Greenwood and Horacio Saggion. A Pattern Based Approach to Answering Factoid, List and Definition Questions. In *Proceedings of the 7th RIAO Conference (RIAO 2004)*, Avignon, France, April 27 2004.

[60] R. Grishman. Information extraction: techniques and challenges. In Maria Teresa Pazienza, editor, *Information Extraction. A multidisciplinary approach to an Emerging Information Technology*, number 1299 in Lecture Notes in Artificial Intelligence. Springer, 1997.

[61] Udo Hahn. Topic Parsing: Accounting for Text Macro Structures in Full-Text Analysis. *Information Processing & Management*, 26(1):135–170, 1990.

[62] Udo Hahn and U. Reimer. Knowledge-Based Text Summarization: Salience and Generalization Operators for Knowledge Base Abstraction. In I. Mani and M.T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 215–232. The MIT Press, 1999.

[63] M.A.K. Halliday and R. Hasan. *Cohesion in English*. Longman Group Limited, 1976.

[64] Sanda Harabagiu, Dan Moldovan, Marius Paşca, Rada Mihalcea, Mihai Surdeanu, Răzvan Bunescu, Roxana Gîrju, Vasile Rus, and Paul Morărescu. FALCON: Boosting Knowledge for Answer Engines. In *Proceedings of the 9th Text REtrieval Conference*, 2000.

[65] Sanda Harabagiu, Dan Moldovan, Marius Paşca, Mihai Surdeanu, Rada Mihalcea, Roxana Gîrju, Vasile Rus, Finley Lăcătuşu, Paul Morărescu, and Răzvan Bunescu. Answering complex, list and context questions with LCC's Question-Answering Server. In *Proceedings of the 10th Text REtrieval Conference*, 2001.

[66] Marti A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of COLING'92*, Nantes, 1992.

[67] Lynette Hirschman and Robert Gaizauskas. Natural Language Question Answering: The View From Here. *Natural Language Engineering*, 7(4), 2001.

[68] J. Hobbs, B. Endres-Niggemeyer, and K. Sparck-Jones, editors. *Summarizing Text for Intelligent Communication*, Dagstuhl,Germany, 1993.

[69] E. Hovy and C-Y. Lin. Automated Text Summarization in SUMMARIST. In I. Mani and M.T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 81–94. The MIT Press, 1999.

[70] Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Michael Junk, and Chin-Yew Lin. Question Answering in Webclopedia. In *Proceedings of the 9th Text REtrieval Conference*, 2000.

[71] Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. Towards Semantics-Based Answer Pinpointing. In *Proceedings of the DARPA Human Language Technology Conference (HLT)*, San Diego, CA, 2001.

[72] Kevin Humphreys, Robert Gaizauskas, Mark Hepple, and Mark Sanderson. University of Sheffield TREC-8 Q & A System. In *Proceedings of the 8th Text REtrieval Conference*, 1999.

[73] John Hutchins. Summarization: Some Problems and Methods. In K.P. Jones, editor, *Meaning: The Frontier of Informatics*, volume 9, pages 151–173. Aslib, 1987.

[74] Abraham Ittycheriah, Martin Franz, Wei-Jing Zhu, Adwait Ratnaparkhi, and Richard J. Mammone. IBM's Statistical Question Answering System. In *Proceedings of the 9th Text REtrieval Conference*, 2000.

[75] D. Jang and S.H. Myaeng. Development of a document summarization system for effective information services. In *RIAO-97. Computer-Assisted Information Searching on Internet.*, pages 101–111, 25th-27th June 1997.

[76] Hongyan Jing. Sentence Reduction for Automatic Text Summarization. In *Proceedings of the 6th Applied Natural Language Processing Conference*, pages 310–315, Seattle, Washington, USA, April 29 - May 4 2000.

[77] Hongyan Jing and Kathleen McKeown. The Decomposition of Human-Written Summary Sentences. In M. Hearst, Gey. F., and R. Tong, editors, *Proceedings of SIGIR'99. 22nd International Conference on Research and Development in Information Retrieval*, pages 129–136, University of California, Beekely, August 1999.

[78] Hongyan Jing and Kathleen McKeown. Cut and Paste Based Text Summarization. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 178–185, Seattle, Washington, USA, April 29 - May 4 2000.

[79] Hongyan Jing, Kathleen McKeown, Regina Barzilay, and Michael Elhadad. Summarization Evaluation Methods: Experiments and Analysis. In *Intelligent Text Summarization. Papers from the 1998 AAAI Spring Symposium. Technical Report SS-98-06*, pages 60–68, Standford (CA), USA, March 23-25 1998. The AAAI Press.

[80] Frances Johnson. Automatic abstracting research. *Library Review*, 44(8):28–36, 1995.

[81] H. Joho and M. Sanderson. Retrieving Descriptive Phrases from Large Amounts of Free Text. In *Proceedings of Conference on Information and Knoweldge Management (CIKM)*, pages 180–186. ACM, 2000.

[82] Paul A. Jones and Chris D. Paice. A 'select and generate' approach to to automatic abstracting. In A.M. McEnry and C.D. Paice, editors, *Proceedings of the 14th British Computer Society Information Retrieval Colloquium*, pages 151–154. Springer Verlag, 1992.

[83] D.E. Kieras. A model of reader strategy for abstracting main ideas from simple technical prose. *Text*, 2(1-3):47–81, 1982.

[84] Walter Kintsch and Teun A. van Dijk. Comment on se rappelle et on résume des histoires. *Langages*, 40:98–116, Décembre 1975.

[85] Kevin Knight and Daniel Marcu. Statistics-based summarization - step one: Sentence compression. In *Proceedings of the 17th National Conference of the American Association for Artificial Intelligence*. AAAI, July 30 - August 3 2000.

[86] Julian Kupiec. MURAX: A Robust Linguistic Approach for Question Answering Using an On-Line Encyclopedia. In *Research and Development in Information Retrieval*, pages 181–190, 1993.

[87] Julian Kupiec, Jan Pedersen, and Francine Chen. A Trainable Document Summarizer. In *Proc. of the 18th ACM-SIGIR Conference*, pages 68–73, 1995.

[88] Cody Kwok, Oren Etzioni, and Daniel S. Weld. Scaling Question Answering to the Web. *ACM Transactions in Information Systems*, 19(3):242–262, July 2001.

[89] F. Lacatusu, L. Hick, S. Harabagiu, and L. Nezd. Lite-GISTexter at DUC2004. In *Proceedings of DUC 2004*. NIST, 2004.

[90] M. Lapata. Probabilistic Text Structuring: Experiments with Sentence Ordering. In *Proceeesingd of the 41st Meeting of the Association of Computational Linguistics*, pages 545–552, Sapporo, 2003.

[91] Abderrafih Lehmam. Une Structuration de Texte Conduisant à la Construction d'un Système de Résumé Automatique. In *Actes des Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'AUPELF-UREF*, pages 175–182, 15 - 16 avril 1997.

[92] Wendy G. Lehnert. Plot Units and Narrative Summarization. *Cognitive Science*, (4):293–331, 1981.

[93] Alessandro Lenci, Roberto Bartolini, Nicoletta Calzolari, Ana Agua, Stephan Busemann, Emmanuel Cartier, Karine Chevreau, and Jos Coch. Multilingual Summarization by Integrating Linguistic Resources in the MLIS-MUSI Project. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02), May 29-31*, Las Palmas, Canary Islands, Spain, 2002.

[94] Xin Li and Dan Roth. Learning Question Classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, 2002.

[95] Elizabeth D. Liddy. The Discourse-Level Structure of Empirical Abstracts: An Exploratory Study. *Information Processing & Management*, 27(1):55–81, 1991.

[96] C. Lin and E. Hovy. Identifying Topics by Position. In *Fifth Conference on Applied Natural Language Processing*, pages 283–290. Association for Computational Linguistics, 31 March - 3 April 1997.

[97] C-Y. Lin. Knowledge-Based Automatic Topic Identification. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. 26-30 June 1995, MIT, Cambridge, Massachusetts, USA*, pages 308–310. ACL, 1995.

[98] C-Y. Lin. Assembly of Topic Extraction Modules in SUMMARIST. In *Intelligent Text Summarization. Papers from the 1998 AAAI Spring Symposium. Technical Report SS-98-06*, pages 53–59, Standford (CA), USA, March 23-25 1998. The AAAI Press.

[99] Dekang Lin and Patrick Pantel. Discovery of Inference Rules for Question Answering. *Natural Language Engineering*, 7(4), 2001.

[100] Jimmy Lin, Dennis Quan, Vineet Sinha, Karun Bakshi, David Huynh, Boris Katz, and David R. Karger. What Makes a Good Answer? The Role of Context in Question Answering. In *Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction (INTERACT 2003)*, Zurich, Switzerland, September 2003.

[101] Lin.C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization*, Barcelona, 2004. ACL.

[102] Hans P. Luhn. The Automatic Creation of Literature Abstracts. *IBM Journal of Research Development*, 2(2):159–165, 1958.

[103] I. Mani and M.T. Maybury, editors. *Advances in Automatic Text Summarization*. The MIT Press, 1999.

[104] Inderjeet Mani. *Automatic Text Summarization*. John Benjamins Publishing Company, 2001.

[105] Inderjeet Mani and Eric Bloedorn. Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1):35–67, 1999.

[106] Inderjeet Mani, Barbara Gates, and Eric Bloedorn. Improving Summaries by Revising Them. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 558–565, Maryland, USA, 20-26 June 1999.

[107] Inderjeet Mani, David House, Gary Klein, Lynette Hirshman, Leo Obrst, Thérèse Firmin, Michael Chrzanowski, and Beth Sundheim. The TIPSTER SUMMAC Text Summarization Evaluation. Technical report, The Mitre Corporation, 1998.

[108] W.C. Mann and S.A. Thompson. Rhetorical Structure Theory: towards a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[109] D. Marcu. *Encyclopedia of Library & Information Science*, chapter Automatic Abstracting, pages 245–256. Miriam Drake, 2003.

[110] Daniel Marcu. From Discourse Structures to Text Summaries. In *The Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 82–88, Madrid, Spain, July 11 1997.

[111] Daniel Marcu. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts.* PhD thesis, Department of Computer Science, University of Toronto, 1997.

[112] Daniel Marcu. To Build Text Summaries of High Quality, Nuclearity is not Sufficient. In *Intelligent Text Summarization*, pages 1–8, Standford (CA), USA, March 23-25 1998.

[113] Daniel Marcu. The automatic construction of large-scale corpora for summarization research. In M. Hearst, Gey. F., and R. Tong, editors, *Proceedings of SIGIR'99. 22nd International Conference on Research and Development in Information Retrieval*, pages 137–144, University of California, Beekely, August 1999.

[114] M.T. Maybury. Generating summaries from event data. *Information Processing & Management*, 31(5):735–751, 1995.

[115] K. McKeown, R. Barzilay, J. Chen, D. Eldon, D. Evans, J. Klavans, A. Nenkova, B. Schiffman, and S. Sigelman. Columbia's Newsblaster: New Features and Future Directions. In *NAACL-HLT'03 Demo*, 2003.

[116] Kathleen McKeown, D. Jordan, and Hatzivassiloglou V. Generating patient-specific summaries of on-line literature. In *Intelligent Text Summarization. Papers from the 1998 AAAI Spring Symposium. Technical Report SS-98-06*, pages 34–43, Standford (CA), USA, March 23-25 1998. The AAAI Press.

[117] Kathleen McKeown, Judith Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. Towards multidocument summarization by reformulation: Progress and prospects. In *AAAI/IAAI*, pages 453–460, 1999.

[118] Kathleen R. McKeown and Dragomir R. Radev. Generating summaries of multiple news articles. In *Proceedings, 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82, Seattle, Washington, July 1995.

[119] Kathleen R. McKeown, J. Robin, and K. Kukich. Generating concise natural language summaries. *Information Processing & Management*, 31(5):702–733, 1995.

[120] K.R. McKeown, J. Robin, and K. Kukich. Generating concise natural language summaries. *Information Processing & Management*, 31(5):702–733, 1995.

[121] S. Miike, E. Itoh, K. Ono, and K. Sumita. A Full-text Retrieval System with A Dynamic Abstract Generation Function. In W.B. Croft and C.J. van Rijsbergen, editors, *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 152–161, July 3-6, Dublin, Ireland, July 3-6 1994.

[122] J-L. Minel, J-P. Desclés, E. Cartier, G. Crispino, S.B. Hazez, and A. Jackiewicz. Résumé automatique par filtrage sémantique d'informations dans des textes. *TSI*, X(X/2000):1–23, 2000.

[123] J-L. Minel, S. Nugier, and G. Piat. Comment Apprécier la Qualité des Résumés Automatiques de Textes? Les Exemples des Protocoles FAN et MLUCE et leurs Résultats sur SERAPHIN. In *1éres Journées Scientificques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'AUPELF-UREF.*, pages 227–232, 15-16 avril 1997.

[124] Ruslan Mitkov. Anaphora resolution: The State of the Art. Working paper, University of Wolverhampton, Wolverhampton, UK, 1999.

[125] Dan Moldovan, Sanda Harabagiu, Roxana Girju, Paul Morărescu, Finley Lăcătuşu, Adrian Novischi, Adriana Badulescu, and Orest Bolohan. LCC Tools for Question Answering. In *Proceedings of the 11th Text REtrieval Conference*, 2002.

[126] Dan Moldovan, Sanda Harabagiu, Marius Paşca, Rada Mihalcea, Richard Goodrum, Roxana Gîrju, and Vasile Rus. The Structure and Performance of an Open-Domain Question Answering System. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 563–570, 2000.

[127] Dan Moldovan, Marius Paşca, Sanda Harabagiu, and Mihai Surdeanu. Performance Issues and Error Analysis in an Open-Domain Question Answering System. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 33–40, Pennsylvania, 2002.

[128] Diego Mollá Aliod, Jawad Berri, and Michael Hess. A Real World Implementation of Answer Extraction. In *Proceedings of the 9th International Conference on Database and Expert Systems Applications Workshop "Natural Language and Information Systems" (NLIS'98)*, Vienna, 1998.

[129] Christof Monz. *From Document Retrieval to Question Answering*. PhD thesis, Institute for Logic, Language and Computation, University of Amsterdam, 2003. Available, April 2004, from `http://www.illc.uva.nl/Publications/Dissertations/DS-2003-04.text.pdf`.

[130] Christof Monz. Minimal Span Weighting Retrieval for Question Answering. In *Proceedings of the SIGIR Workshop on Information Retrieval for Question Answering (IR4QA)*, Sheffield, UK, July 29 2004.

[131] Ani Nenkova and Rebecca Passonneau. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of NAACL-HLT 2004*, 2004.

[132] Michael P. Oakes and Chris D. Paice. The Automatic Generation of Templates for Automatic Abstracting. In *21st BCS IRSG Colloquium on IR*, Glasgow, 1999.

[133] Michael P. Oakes and Chris D. Paice. Term extraction for automatic abstracting. In D. Bourigault, C. Jacquemin, and M-C. L'Homme, editors, *Recent Advances in Computational Terminology*, volume 2 of *Natural Language Processing*, chapter 17, pages 353–370. John Benjamins Publishing Company, 2001.

[134] Kenji Ono, Kazuo Sumita, and Seiji Miike. Abstract Generation Based on Rhetorical Structure Extraction. In *Proceedings of the International Conference on Computational Linguistics*, pages 344–348, 1994.

[135] Chris D. Paice. The Automatic Generation of Literary Abtracts: An Approach based on Identification of Self-indicating Phrases. In O.R. Norman, S.E. Robertson, C.J. van Rijsbergen, and P.W. Williams, editors, *Information Retrieval Research*, London: Butterworth, 1981.

[136] Chris D. Paice. Constructing Literature Abstracts by Computer: Technics and Prospects. *Information Processing & Management*, 26(1):171–186, 1990.

[137] Chris D. Paice, William J. Black, Frances C. Johnson, and A.P. Neal. Automatic Abstracting. Technical Report R&D Report 6166, British Library, 1994.

[138] Chris D. Paice and Paul A. Jones. The Identification of Important Concepts in Highly Structured Technical Papers. In R. Korfhage, E. Rasmussen, and P. Willett, editors, *Proc. of the 16th ACM-SIGIR Conference*, pages 69–78, 1993.

[139] Chris D. Paice and Michael P. Oakes. A Concept-Based Method for Automatic Abstracting. Technical Report Research Report 27, Library and Information Commission, 1999.

[140] K. Pastra and H. Saggion. Colouring summaries Bleu. In *Proceedings of Evaluation Initiatives in Natural Language Processing*, Budapest, Hungary, 14 April 2003. EACL.

[141] Jennifer Pearson. *Terms in Context*, volume 1 of *Studies in Corpus Linguistics*. Jhon Benjamins Publishing Company, 1998.

[142] M. Pinto Molina. Documentary Abstracting: Towards a Methodological Model. *Journal of the American Society for Information Science*, 46(3):225–234, April 1995.

[143] Luc Plamondon, Guy Lapalme, and Leila Kosseim. The QUANTUM Question Answering System. In *Proceedings of the 10th Text REtrieval Conference*, 2001.

[144] J.J. Pollock and A. Zamora. Automatic abstracting research at Chemical Abstracts Service. *Journal of Chemical Information and Computer Sciences*, (15):226–233, 1975.

[145] Hong Qi, Jahna Otterbacher, Adam Winkel, and Dragomir R. Radev. The University of Michigan at TREC2002: Question Answering and Novelty Tracks. In *Proceedings of the 11th Text REtrieval Conference*, 2002.

[146] Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *ANLP/NAACL Workshop on Summarization*, Seattle, WA, April 2000.

[147] Dragomir R. Radev and Kathleen R. McKeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500, September 1998.

[148] Lisa F. Rau and Ronald Brandow. Domain-independent summarization of news. Dagstuhl Seminar, Summarizing Text for Intelligent Communication, December 1993.

[149] Lisa F. Rau, Paul S. Jacobs, and Uri Zernik. Information Extraction and Text Summarization using Linguistic Knowledge Acquisition. *Information Processing & Management*, 25(4):419–428, 1989.

[150] Deepak Ravichandran and Eduard Hovy. Learning Surface Text Patterns for a Question Answering System. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 41–47, Pennsylvania, 2002.

[151] J.B. Reiser, J.B. Black, and W. Lehnert. Thematic knowledge structures in the understanding and generation of narratives. *Discourse Processes*, (8):357–389, 1985.

[152] *RIFRA'98. Rencontre Internationale sur l'extraction le Filtrage et le Résumé Automatique. Novembre 11-14*, Sfax, Tunisie, Novembre 11-14 1998.

[153] Lucia H.M. Rino and Donia Scott. Automatic generation of draft summaries: Heuristics for content selection. Technical Report ITRI-94-8, Information Technology Research Institute, 1994.

[154] Lucia H.M. Rino and Donia Scott. A Discourse Model for Gist Preservation. In D.L. Borges and C.A.A. Kaestner, editors, *Proceedings of the 13th Brazilian Symposium on Artificial Intelligence, SBIA'96*, Advances in Artificial Intelligence, pages 131–140. Springer, October 23-25, Curitiba, Brazil 1996.

[155] Ian Roberts and Robert Gaizauskas. Evaluating Passage Retrieval Approaches for Question Answering. In *Proceedings of 26th European Conference on Information Retrieval*, 2004.

[156] Richard Robinson. *Definition*. Oxford University Press, 1954.

[157] Jennifer Rowley. *Abstracting and Indexing*. Clive Bingley, London, 1982.

[158] J.E. Rush, R. Salvador, and A. Zamora. Automatic Abstracting and Indexing. Production of Indicative Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria. *Journal of the American Society for Information Science*, pages 260–274, July-August 1971.

[159] H. Saggion. Shallow-based Robust Summarization. In *Automatic Summarization: Solutions and Perspectives*, ATALA, December, 14 2002.

[160] H. Saggion and R. Gaizauskas. Mining on-line sources for definition knowledge. In *Proceedings of the 17th FLAIRS 2004*, Miami Bearch, Florida, USA, May 17-19 2004. AAAI.

[161] H. Saggion and G. Lapalme. Concept Identification and Presentation in the Context of Technical Text Summarization. In *Proceedings of the Workshop on Automatic Summarization. ANLP-NAACL2000*, Seattle, WA, USA, 30 April 2000. Association for Computational Linguistics.

[162] H. Saggion and G. Lapalme. Generating Indicative-Informative Summaries with SumUM. *Computational Linguistics*, 2002.

[163] H. Saggion, D. Radev, S. Teufel, and W. Lam. Meta-evaluation of Summaries in a Cross-lingual Environment using Content-based Metrics. In *Proceedings of COLING 2002*, pages 849–855, Taipei, taiwan, August 24 - September 1 2002.

[164] H. Saggion, D. Radev, S. Teufel, L. Wai, and S. Strassel. Developing Infrastructure for the Evaluation of Single and Multi-document Summarization Systems in a Cross-lingual Environment. In *LREC 2002*, pages 747–754, Las Palmas, Gran Canaria, Spain, 2002.

[165] Horacio Saggion and Robert Gaizauskas. Multi-document summarization by cluster/profile relevance and redundancy removal. In *Proceedings of the Document Understanding Conference 2004*. NIST, 2004.

[166] Horacio Saggion, Robert Gaizauskas, Mark Hepple, Ian Roberts, and Mark A. Greenwood. Exploring the Performance of Boolean Retrieval Strategies for Open Domain Question Answering. In *Proceedings of the SIGIR Workshop on Information Retrieval for Question Answering (IR4QA)*, Sheffield, UK, July 29 2004.

[167] Horacio Saggion and Guy Lapalme. Where does Information come from? Corpus Analysis for Automatic Abstracting. In *Rencontre Internationale sur l'Extraction le Filtrage et le Résumé Automatique. RIFRA'98*, pages 72–83, Sfax, Tunisie, Novembre 11-14 1998.

[168] Saggion, H. and Bontcheva, K. and Cunningham, H. Generic and Query-based Summarization. In *European Conference of the Association for Computational Linguistics (EACL) Research Notes and Demos*, Budapest, Hungary, 12-17 April 2003. EACL.

[169] G. Salton. *Automatic Text Processing*. Addison-Wesley Publishing Company, 1988.

[170] Gerald Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. Automatic Text Structuring and Summarization. *Information Processing & Management*, 33(2):193–207, 1997.

[171] B. Schiffman, I. Mani, and K.J. Concepcion. Producing Biographical Summaries: Combining Linguistic Knowlkedge with Corpus Statistics. In *Proceedings of EACL-ACL*, 2001.

[172] Sam Scott and Robert Gaizauskas. University of Sheffield TREC-9 Q & A System. In *Proceedings of the 9th Text REtrieval Conference*, 2000.

[173] R. Shank and R. Abelson. *Scripts Plans Goals and Understanding*. Lawrence Erlbaum Associates, Publishers, 1977.

[174] Bernadette Sharp. *Elaboration and Testing of New Methodologies for Automatic Abstracting*. PhD thesis, The University of Aston in Birmingham., October 1989.

[175] Gerardo Sierra, Alfonso Medina, Rodrigo Alarcón, and César A. Aguilar. Towards the Extraction of Conceptual Information From Corpora. In D. Archer, P. Rayson, A. Wilson, and T. McEnery, editors, *Proceedings of the Corpus Linguistics 2003 Conference*, pages 691–697. University Centre for Computer Corpus Research on Language, 2003.

[176] Robert F. Simmons. Answering English Questions by Computer: A Survey. *Communications of the ACM*, 8(1):53–70, 1965.

[177] Martin M. Soubbotin and Sergei M. Soubbotin. Patterns of Potential Answer Expressions as Clues to the Right Answers. In *Proceedings of the 10th Text REtrieval Conference*, 2001.

[178] Karen Sparck Jones. Discourse Modelling for Automatic Summarising. Technical Report 290, University of Cambridge, Computer Laboratory, February 1993.

[179] Karen Sparck Jones. What Might Be in a Summary? In K. Knorz and Womser-Hacker, editors, *Information Retrieval 93: Von der Modellierung zur Anwendung*, 1993.

[180] Karen Sparck Jones. Document Processing: Summarization. In R. Cole, editor, *Survey of the State of the Art in Human Language Technology*, chapter 7, pages 266–269. Cambridge University Press, 1997.

[181] Karen Sparck Jones. Automatic Summarizing: Factors and Directions. In I. Mani and M. Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press, Cambridge MA, 1999.

[182] Karen Sparck Jones and Brigitte Endres-Niggemeyer. Automatic Summarizing. *Information Processing & Management*, 31(5):625–630, 1995.

[183] Karen Sparck Jones and Julia R. Galliers. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Number 1083 in Lecture Notes in Artificial Intelligence. Springer, 1995.

[184] T. Strzalkowski, J. Wang, and Wise B. A Robust Practical Text Summarization. In *Intelligent Text Summarization Symposium (Working Notes)*, pages 26–33, Standford (CA), USA, March 23-25 1998.

[185] B. Sundheim, editor. *Proceedings of the Sixth Message Understanding Conference*, Columbia, MD, November 1995. Morgan Kaufman.

[186] John I. Tait. *Automatic Summarising of English Texts*. PhD thesis, University of Cambridge, Computer Laboratory, December 1982.

[187] Simone Teufel. Meta-Discourse Markers and Problem-Structuring in Scientific Texts. In M. Stede, L. Wanner, and E. Hovy, editors, *Proceedings of the Workshop on Discourse Relations and Discourse Markers, COLING-ACL'98*, pages 43–49, 15th August 1998.

[188] Simone Teufel and Marc Moens. Argumentative classification of extracted sentences as a first step towards flexible abstracting. In I. Mani and M.T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 155–171. The MIT Press, 1999.

[189] Translingual Information Detection, Extraction and Summarization (TIDES) Program. http://www.darpa.mil/ito/research/tides/index.html, August 2000.

[190] Anastasios Tombros, Mark Sanderson, and Phil Gray. Advantages of Query Biased Summaries in Information Retrieval. In *Intelligent Text Summarization. Papers from the 1998 AAAI Spring Symposium. Technical Report SS-98-06*, pages 34–43, Standford (CA), USA, March 23-25 1998. The AAAI Press.

[191] Peter D. Turney. Learning to Extract Keyphrases from Text. Technical Report NRC Technical Report ERB-1051, National Research Council of Canada, 1999.

[192] Teun A. van Dijk. Recalling and Summarizing Complex Discourse. In M.A. Just and Carpenters, editors, *Cognitive Processes in Comprehension*, 1977.

[193] C.J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.

[194] Ellen M. Voorhees. The TREC 8 Question Answering Track Report. In *Proceedings of the 8th Text REtrieval Conference*, 1999.

[195] Ellen M. Voorhees. Overview of the TREC-9 Question Answering Track. In *Proceedings of the 9th Text REtrieval Conference*, 2000.

[196] Ellen M. Voorhees. Overview of the TREC 2001 Question Answering Track. In *Proceedings of the 10th Text REtrieval Conference*, 2001.

[197] Ellen M. Voorhees. Overview of the TREC 2002 Question Answering Track. In *Proceedings of the 11th Text REtrieval Conference*, 2002.

[198] Ellen M. Voorhees. Evaluating Answers to Definition Questions. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*, volume 2, pages 109–111, 2003.

[199] Ellen M. Voorhees. Overview of the TREC 2003 Question Answering Track. In *Proceedings of the 12th Text REtrieval Conference*, 2003.

[200] Ellen M. Voorhees. Overview of the TREC 2004 Question Answering Track. In *Proceedings of the 13th Text REtrieval Conference*, 2004.

[201] J. Xu, A. Licuanan, and R. Weischedel. TREC2003 QA at BBN: Answering Definitional Questions. In *Proceedings of TREC-2003*, 2003.

[202] Jinxi Xu, Ralph Weischedel, and Ana Licuanan. Evaluation of an Extraction-Based Approach to Answering Definitional Questions. In *Proceedings of the 27th ACM International Conference on Research and Development in Information Retrieval (SIGIR 2004)*, Sheffield, UK, July 2004.

[203] Hui Yang, Hang Cui, Mstisalv Maslennikov, Long Qiu, Min-Yen Kan, and Tat-Seng Chua. QUALIFIER in TREC-12 QA Main Task. In *Proceedings of the 12th Text REtrieval Conference*, 2003.

[204] Dell Zhang and Wee Sun Lee. Question Classification using Support Vector Machines. In *Proceedings of the 26th ACM International Conference on Research and Developement in Information Retrieval (SIGIR'03)*, Toronto, Canada, 2003.

[205] Zhiping Zheng. AnswerBus Question Answering System. In *Proceedings of the Human Language Technology Conference (HLT 2002)*, San Diego, CA, March 24-27 2002.

[206] L. Zhou, M. Ticrea, and E. Hovy. Multi-document Biography Summarization. In *Proceedings of Empirical Methods in Natural Language Processing*, 2004.