Automatic Text Summarization

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

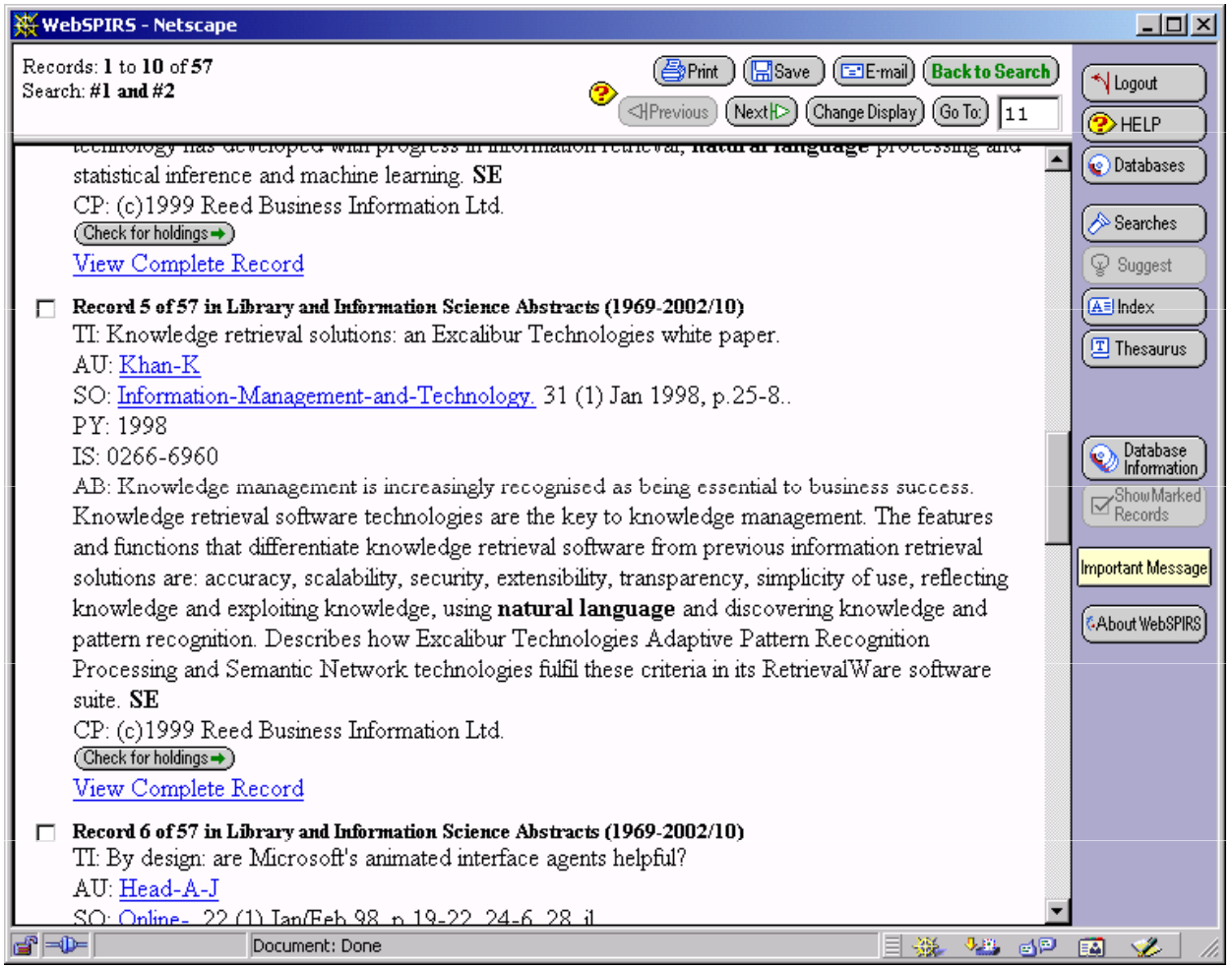
Outline

- Summarization Definitions
- Summary Typology
- Automatic Summarization
- Summarization by Sentence Extraction
- Superficial Features
- Learning Summarization Systems
- Cohesion-based Summarization
- Rhetorical-based Summarization
- Non-extractive Summarization
- Information Extraction and Summarization
- Headline Generation & Cut and Paste Summarization & Paraphrase Generation
- Multi-document Summarization
- Summarization Evaluation
- SUMMAC Evaluation
- DUC Evaluation
- Other Evaluations
- Rouge & Pyramid Metrics
- MEAD System
- SUMMA System
- Summarization Resources

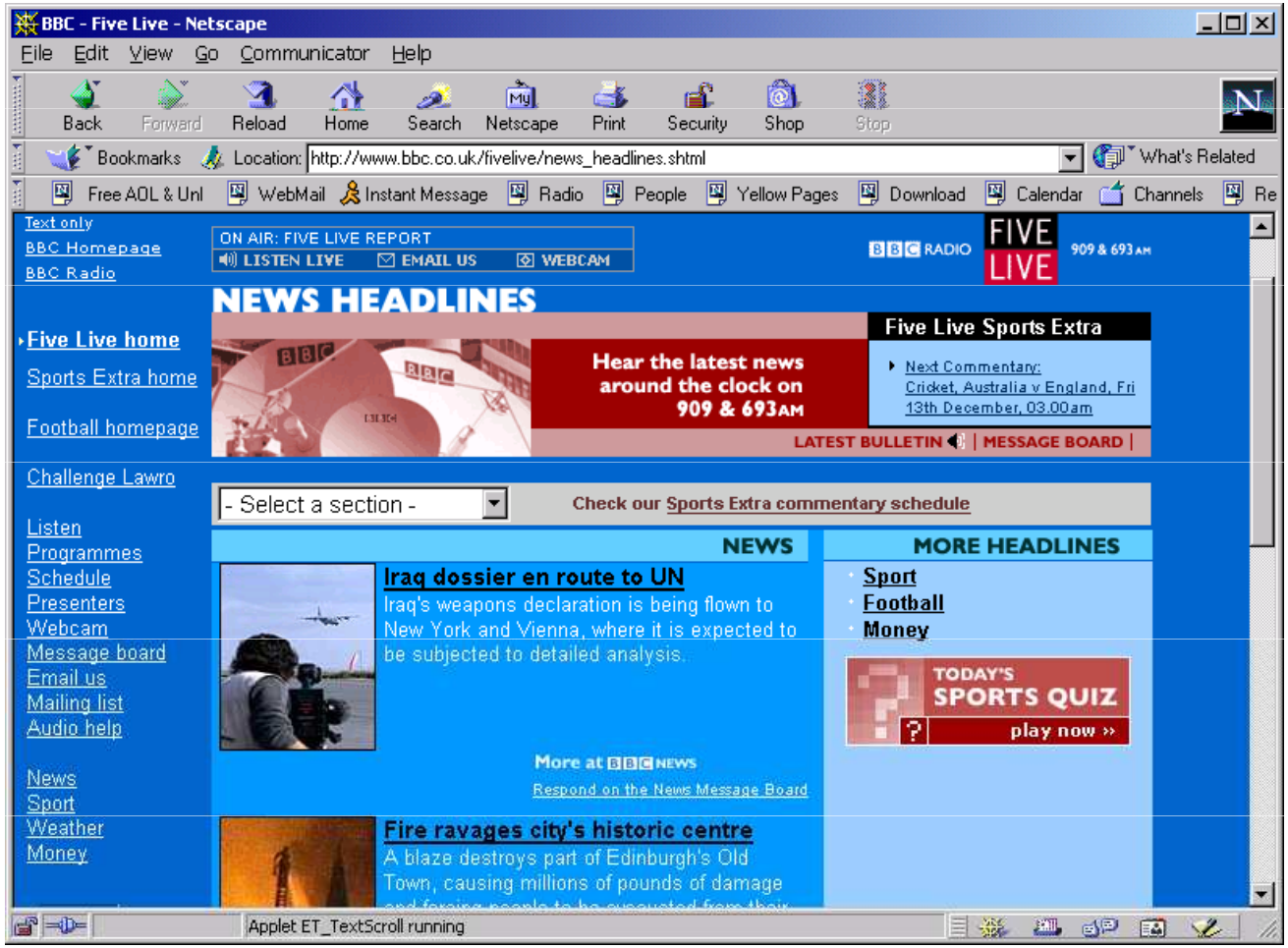
Automatic Text Summarization

- 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

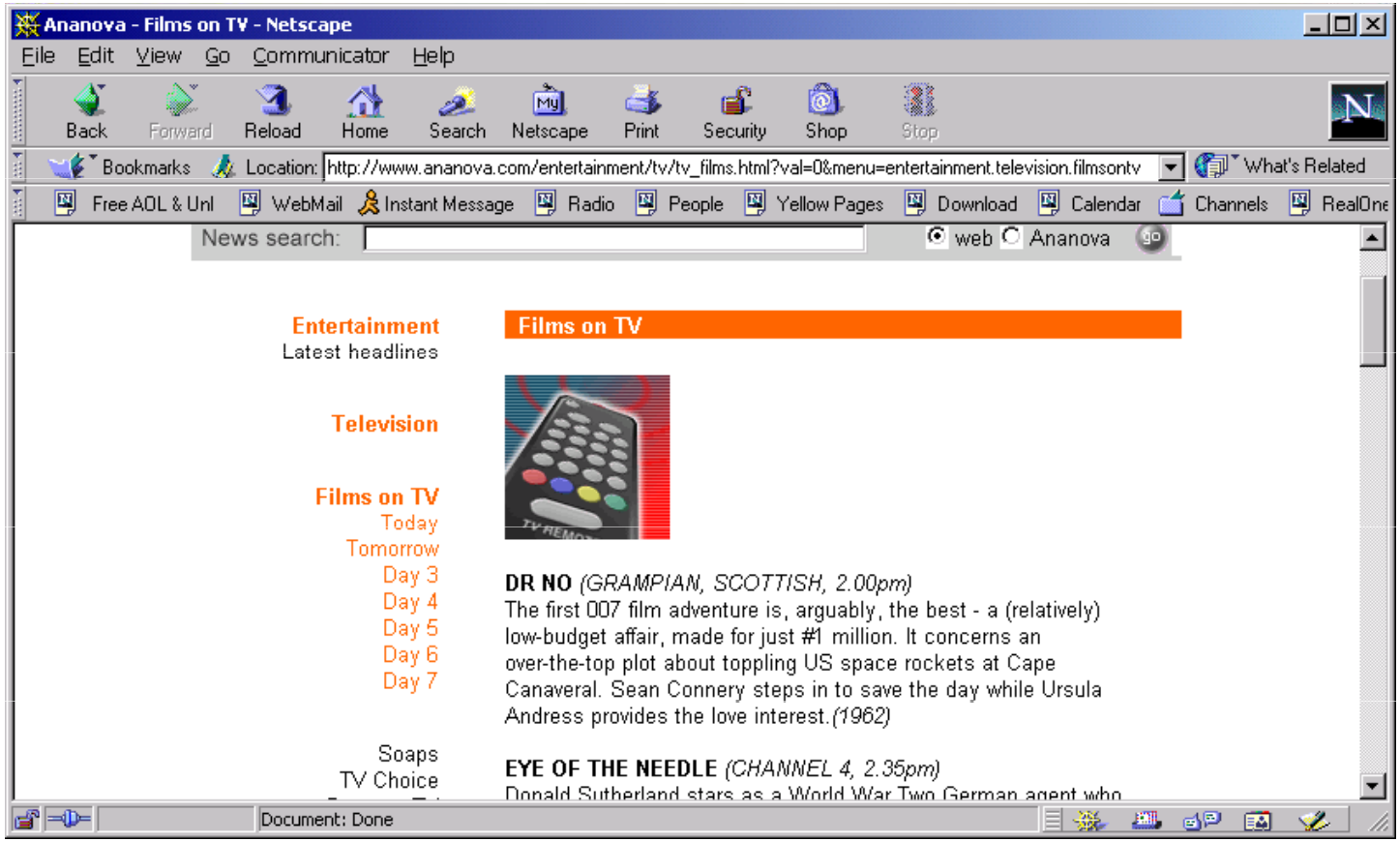
Examples of summaries – abstract of research article



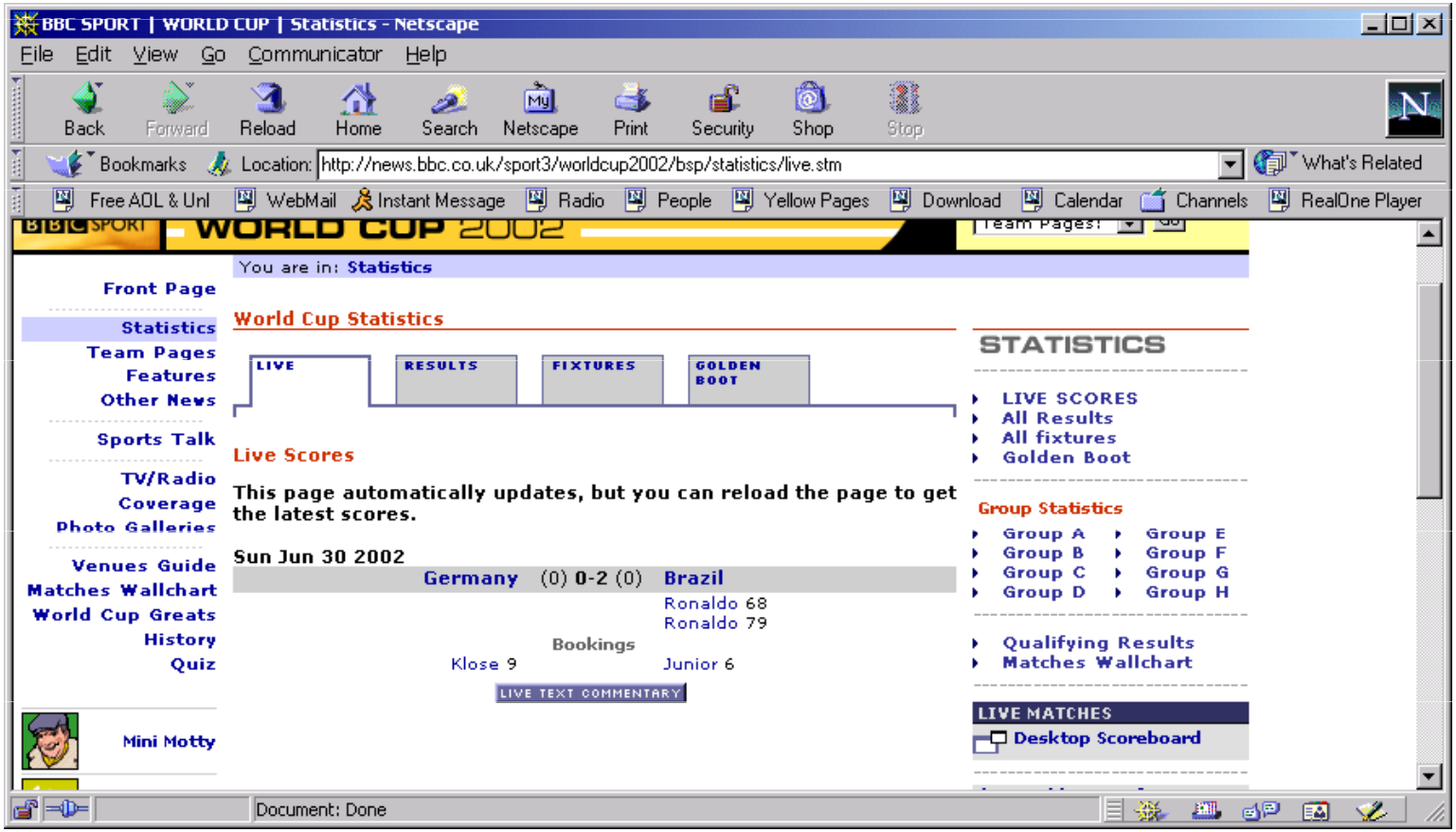
Examples of summaries – headline + leading paragraph



Examples of summaries – movie preview



Examples of summaries – sports results



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 experitise.

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

NLP for 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.

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/...

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 $\frac{TP}{TP + FN}$

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

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

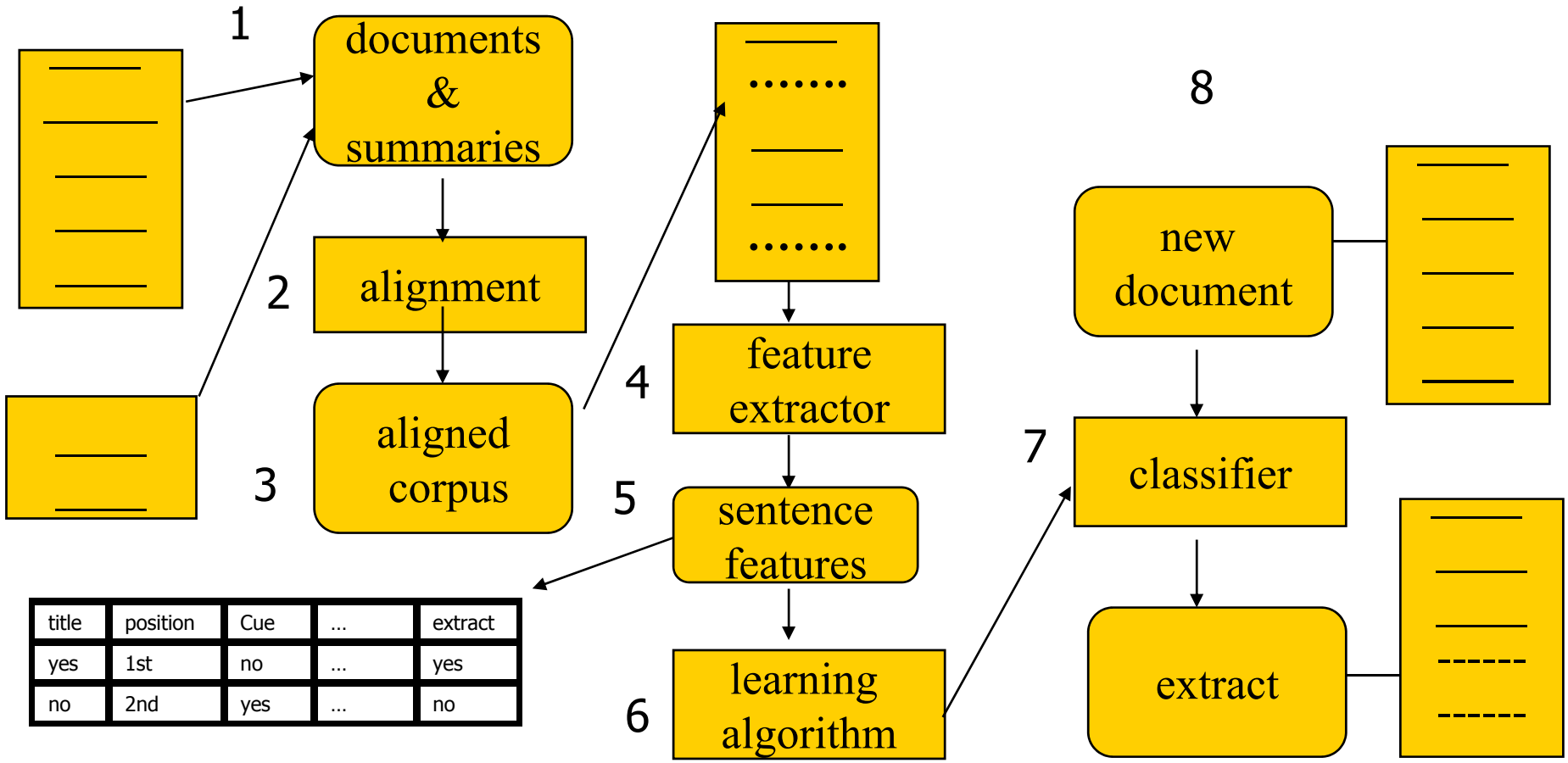
$$Weight(S) = \alpha.Title(S) + \beta.Cue(S) + \gamma.Keyword(S) + \delta.Position(S)$$

  - 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



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

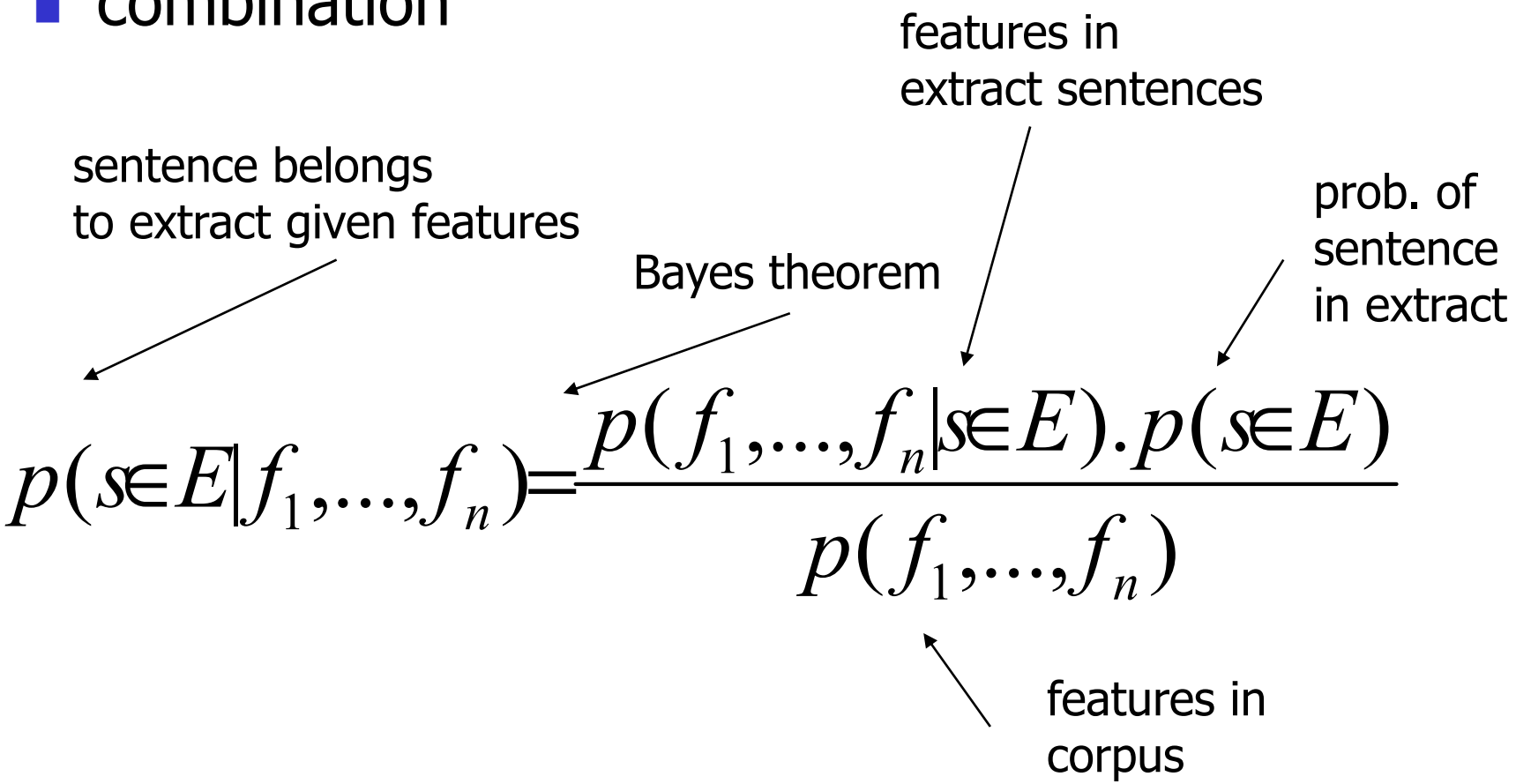
$$heading(S_{i-1}) \wedge (S_{i-1} \cap DIC_{headings}) \neq \phi$$

Statistical combination

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

Statistical combination

- combination



Statistical combination

- parameter estimation

assume independence

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

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

estimate by counting

$$p(s \in E)$$

Statistical combination

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

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.

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.

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

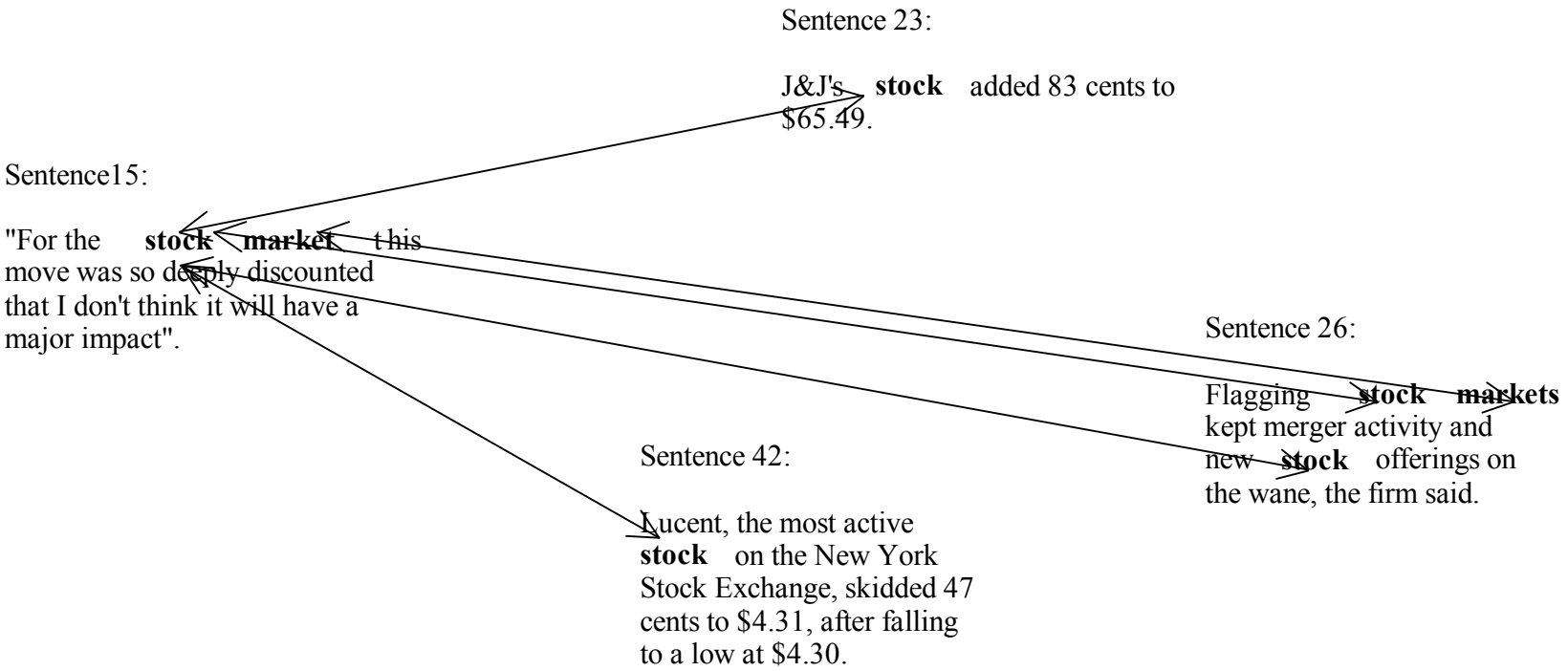
Approaches to cohesion

- 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 10th word of sentence
    - else: internal
  - selection (indicator) & rejection & aggregation rules; reported success: abstract > aggregation > extract

Telepattan system: (Bembrahim & Ahmad'95)

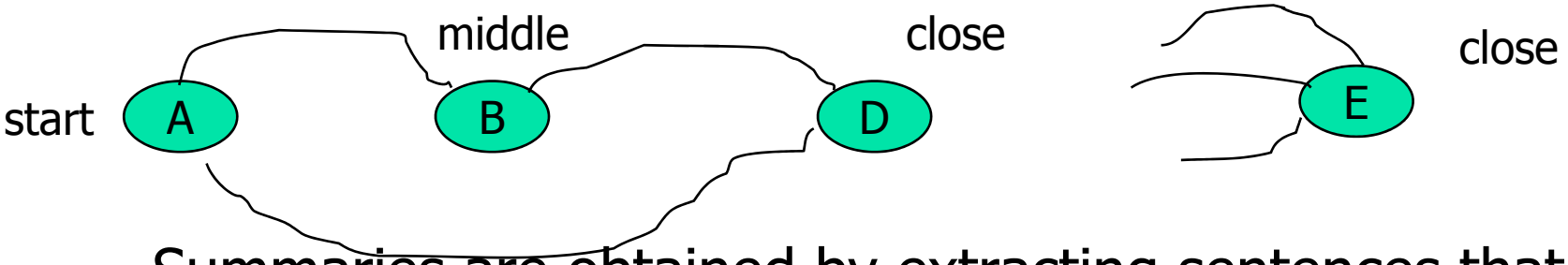
- Link two sentences if
  - they contain words related by repetition, synonymy, class/superclass (hypernymy), paraphrase
    - *destruct* ~ *destruction*
  - use thesaurus (i.e., related words)
- pruning
  - links(s_i, s_j) > thr => bond (s_i, s_j)

Telepattan system



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 to and from a given sentence



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

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

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

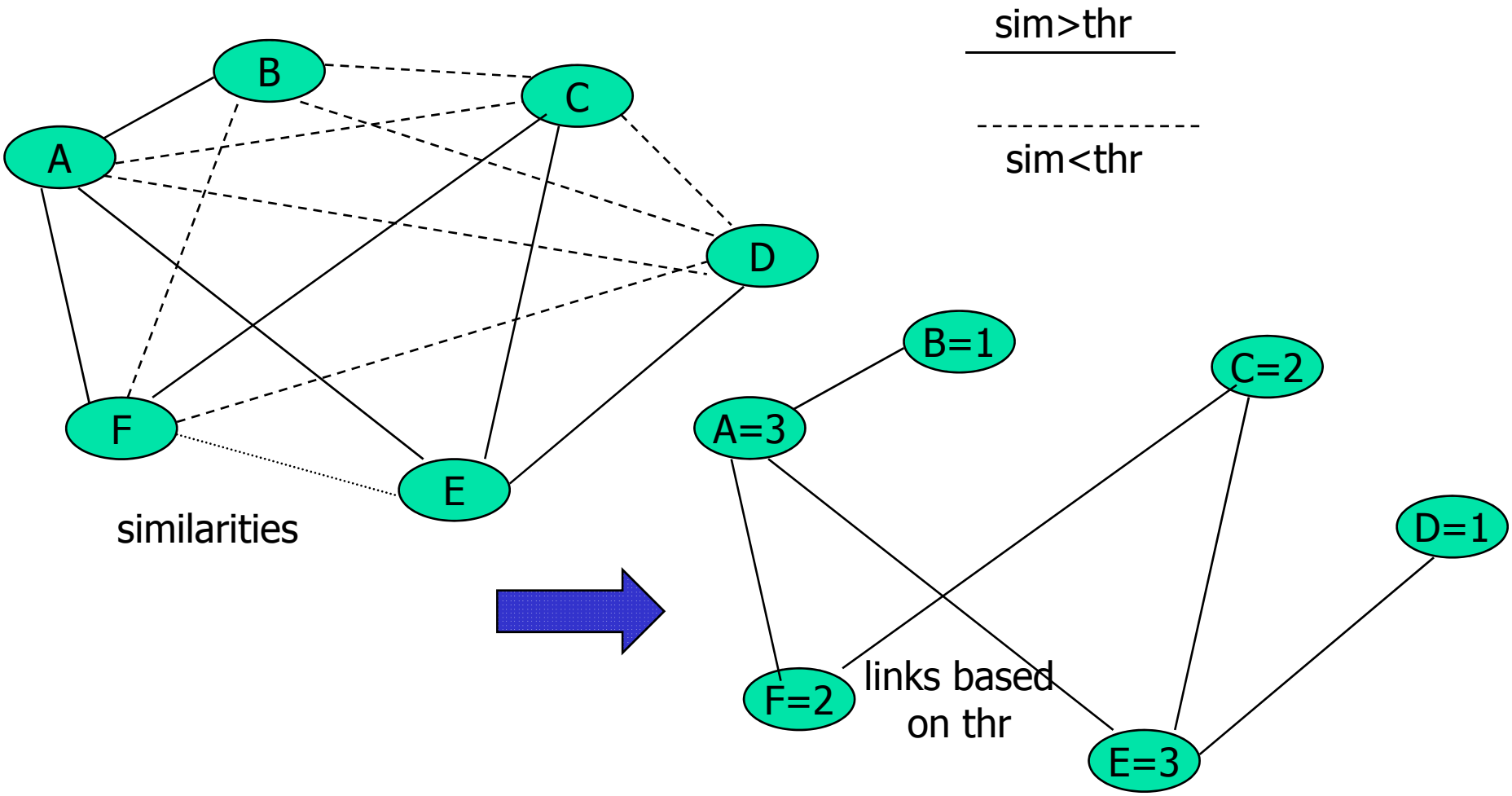
Information retrieval techniques (Salton&al'97)

- 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

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

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."

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

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.

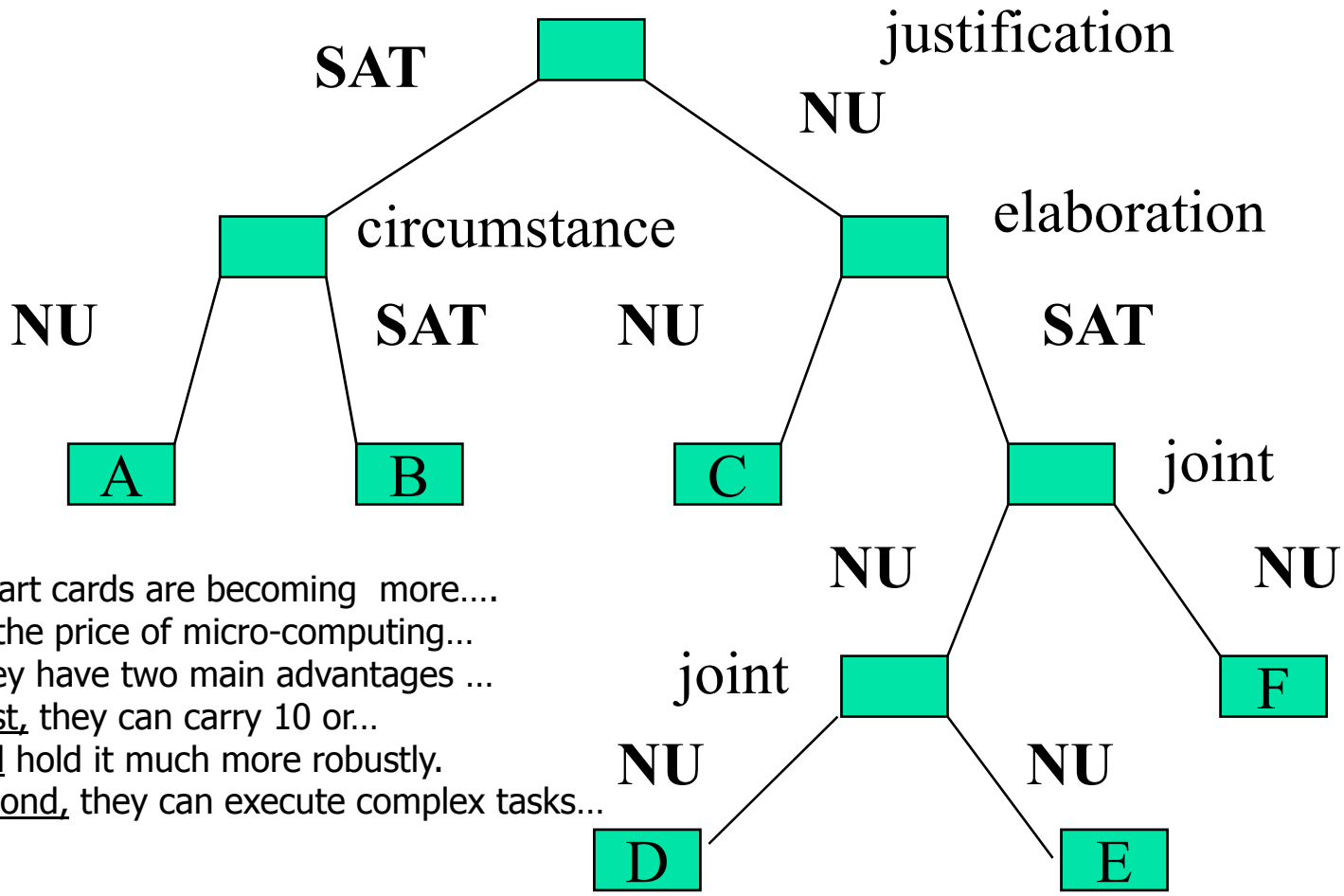
(C) They have two main advantages over magnetic strip cards.

(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.

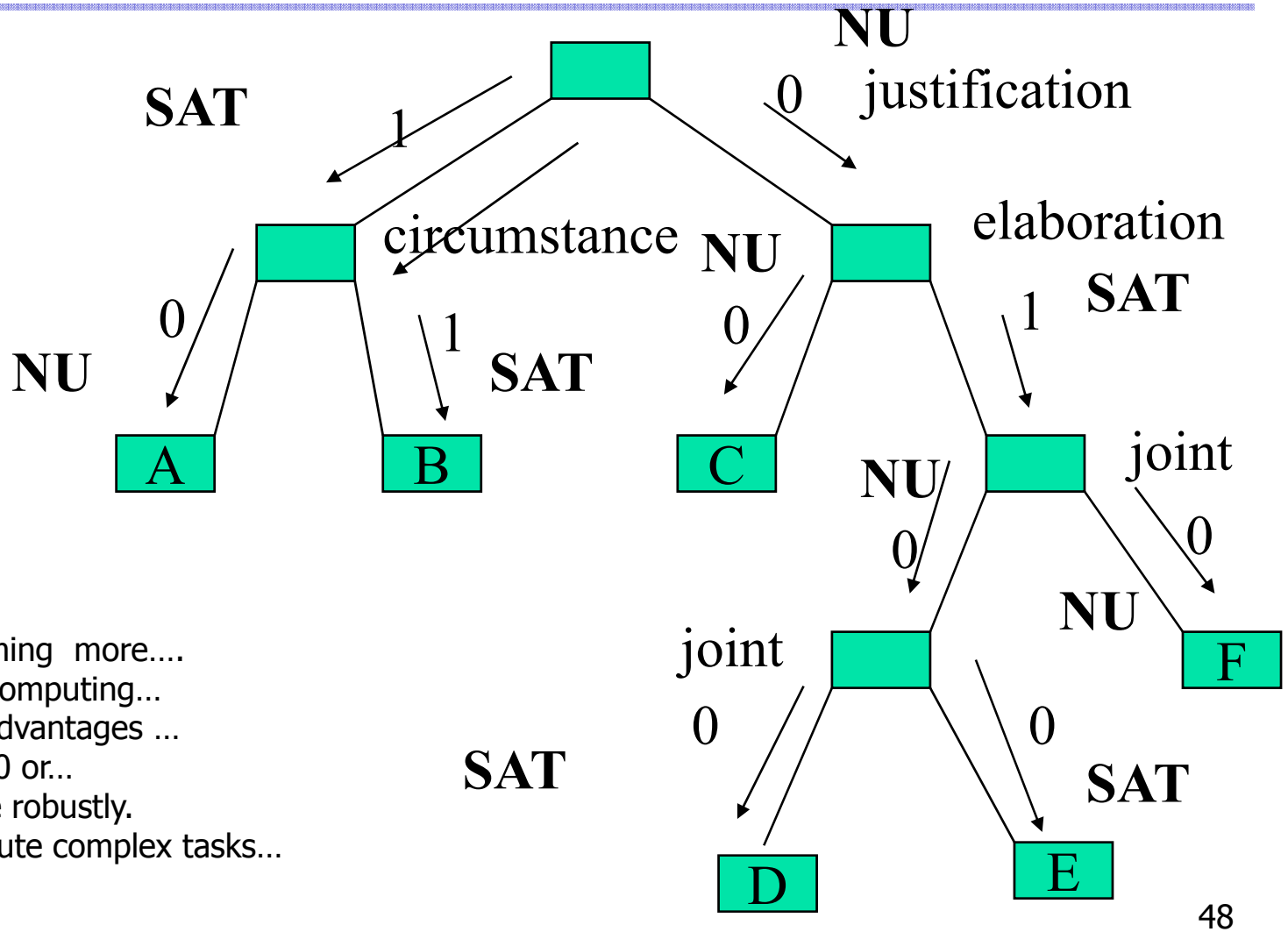
Rhetorical tree



(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…

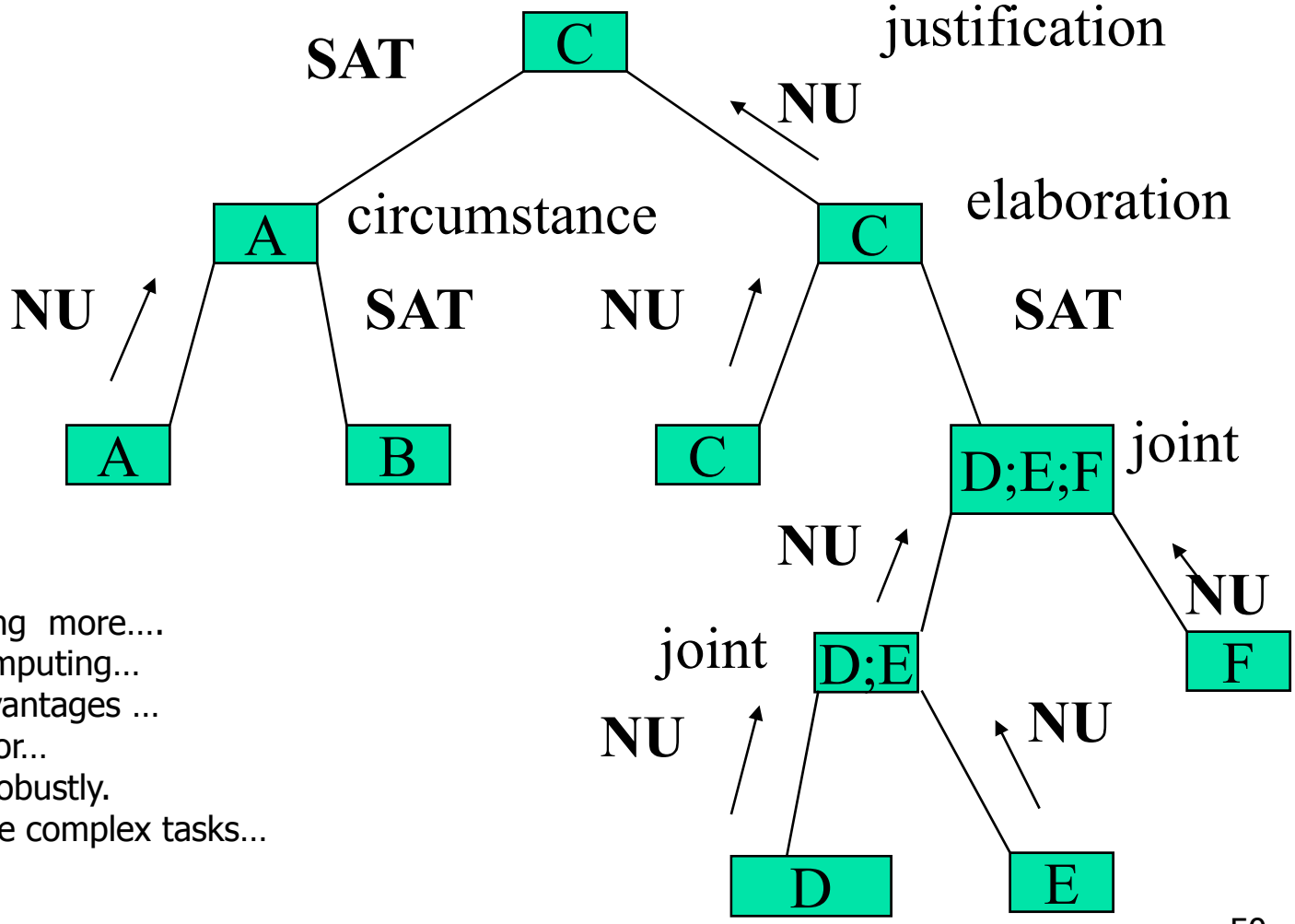
RTS extract

(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.

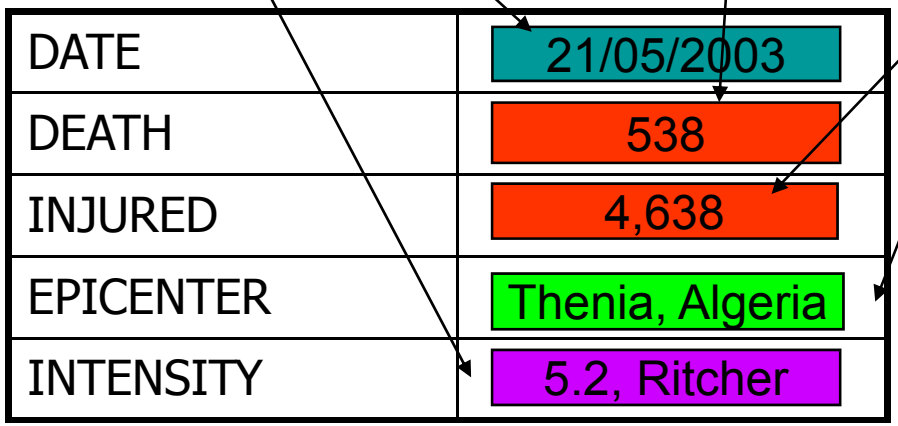
Information Extraction

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

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, ...



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.

There was an earthquake in Illinois with a 3.2 Richter scale.

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

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: $P(I_{k+1}=<S2,W2> \mid I_k=<S1,W1>)$
  - 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

Paraphrase

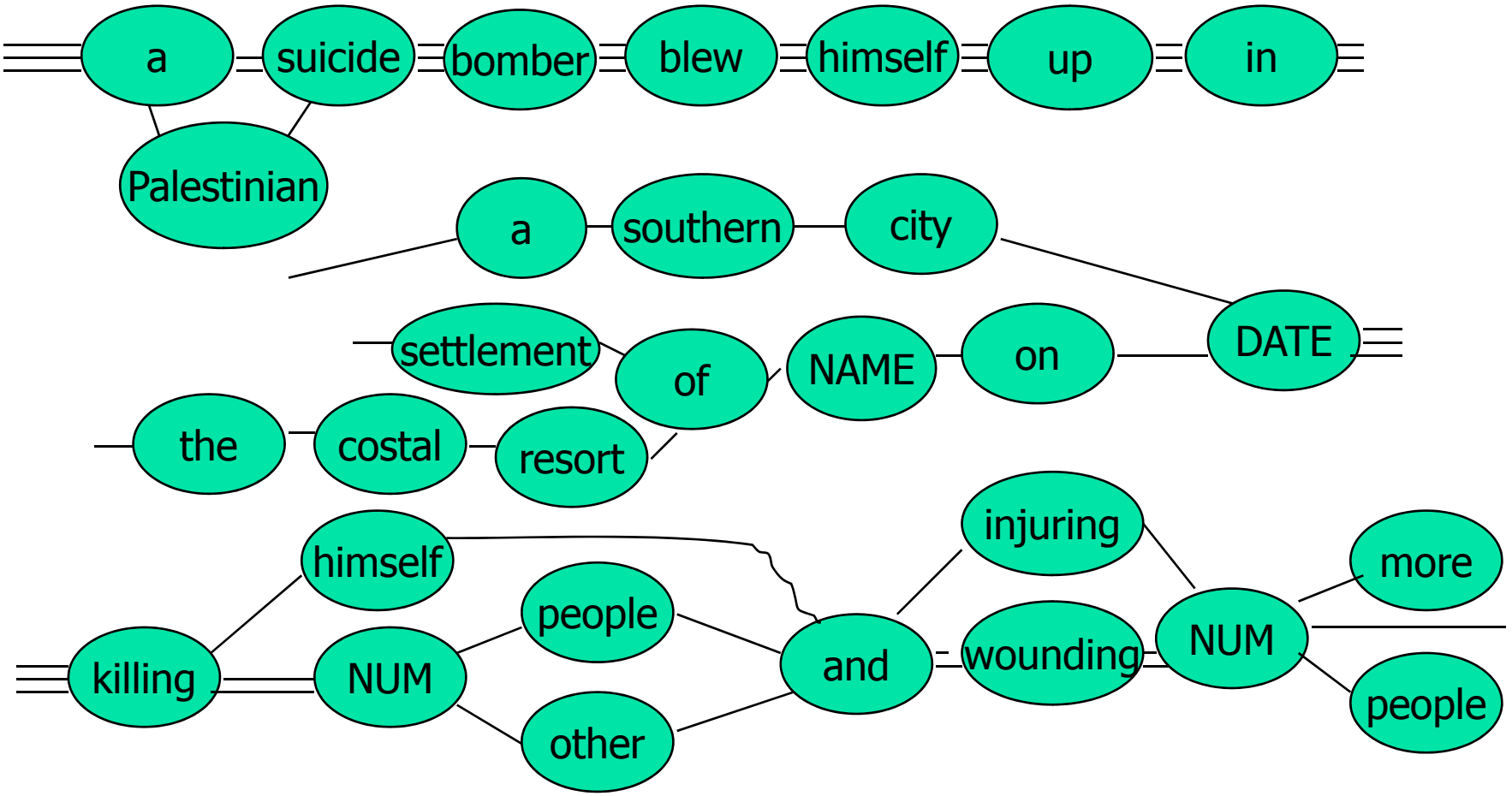
- Alignment based paraphrase: Barzilay&Lee'2003
- unsupervised approach to learn:
  - patterns in the data & equivalences among patterns
  - X injured Y people, Z seriously = Y were injured by X among them Z were in serious condition
  - learning is done over two different corpus which are comparable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

Similar event descriptions

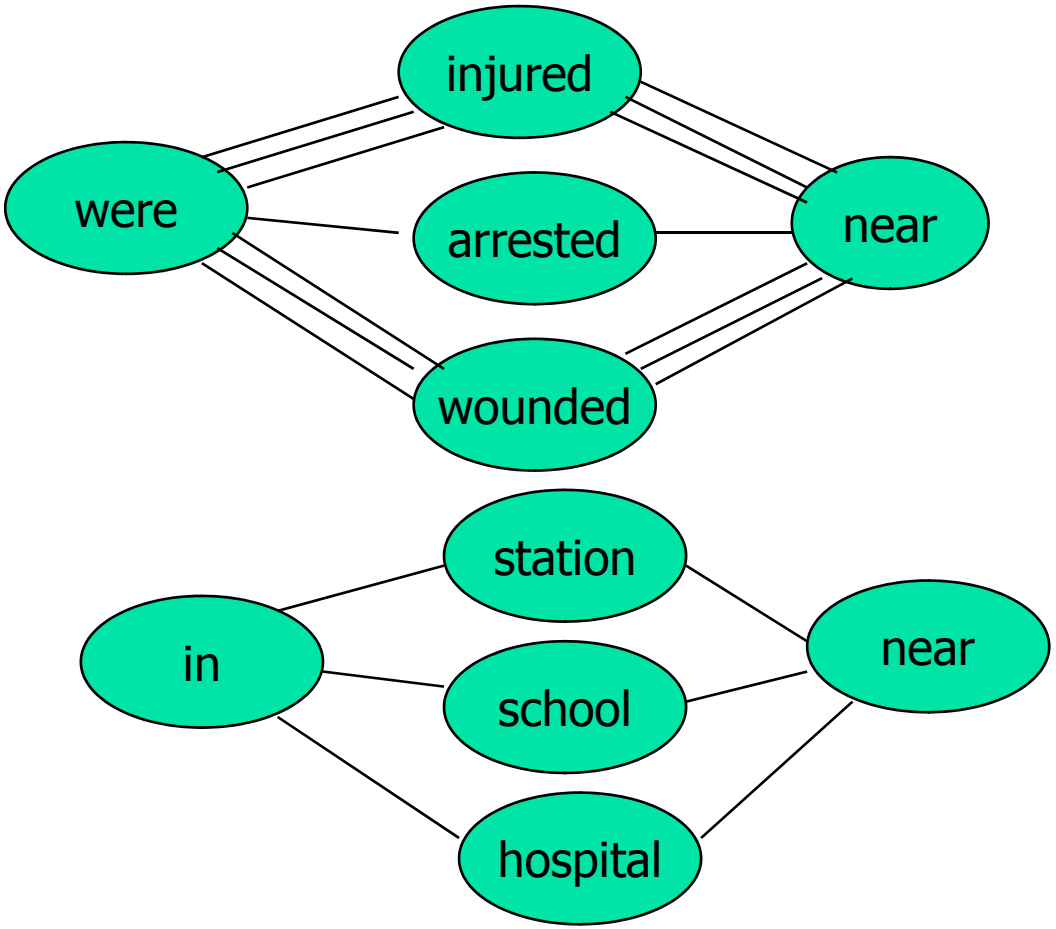
- **Cluster of similar sentences**
  - **A Palestinian suicide bomber blew himself up in** a southern city Wednesday, **killing** two other **people and wounding** 27.
  - **A suicide bomber blew himself up in** the settlement of Efrat, on Sunday, **killing** himself **and injuring** seven people.
  - **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.

Lattices and backbones



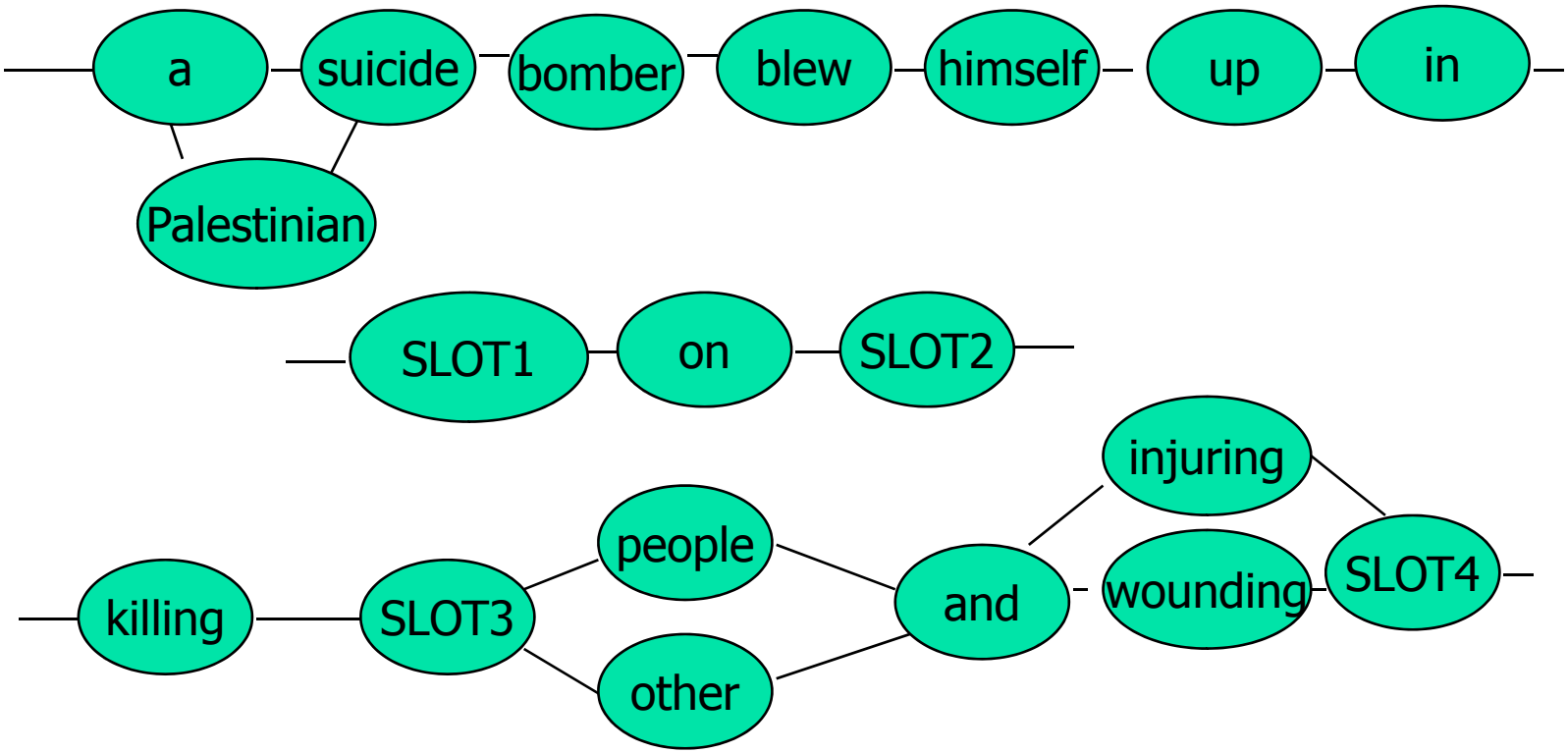
Arguments or Synonyms?



keep words

replace by arguments

Patterns induced



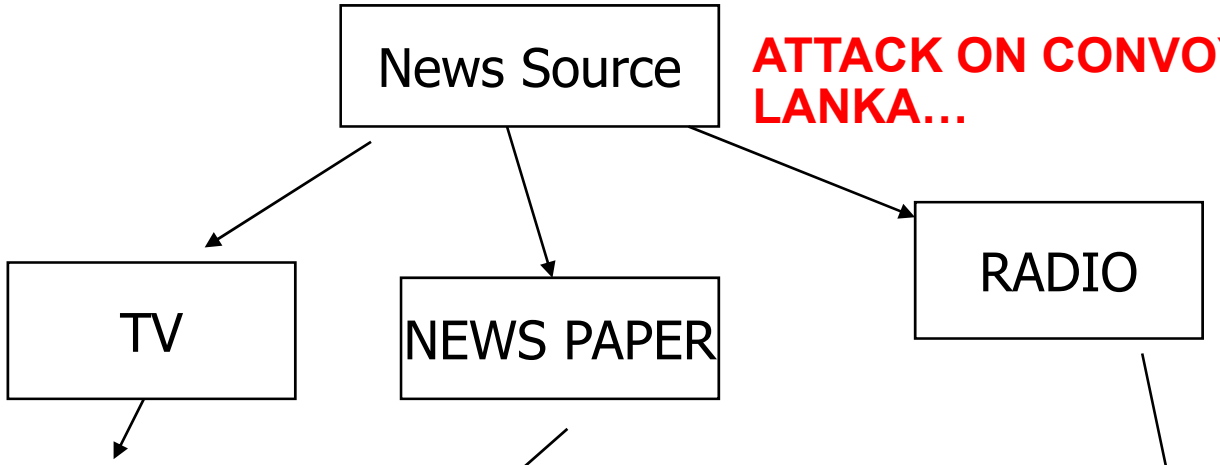
Generating paraphrases

- 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

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



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"

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

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

similarity to document already seen

Q = query
R = list of documents
D_k = k - document in list
S = subset of R already scanned

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.

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) $\frac{TP+TN}{TP+FP+FP+FN}$

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

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

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

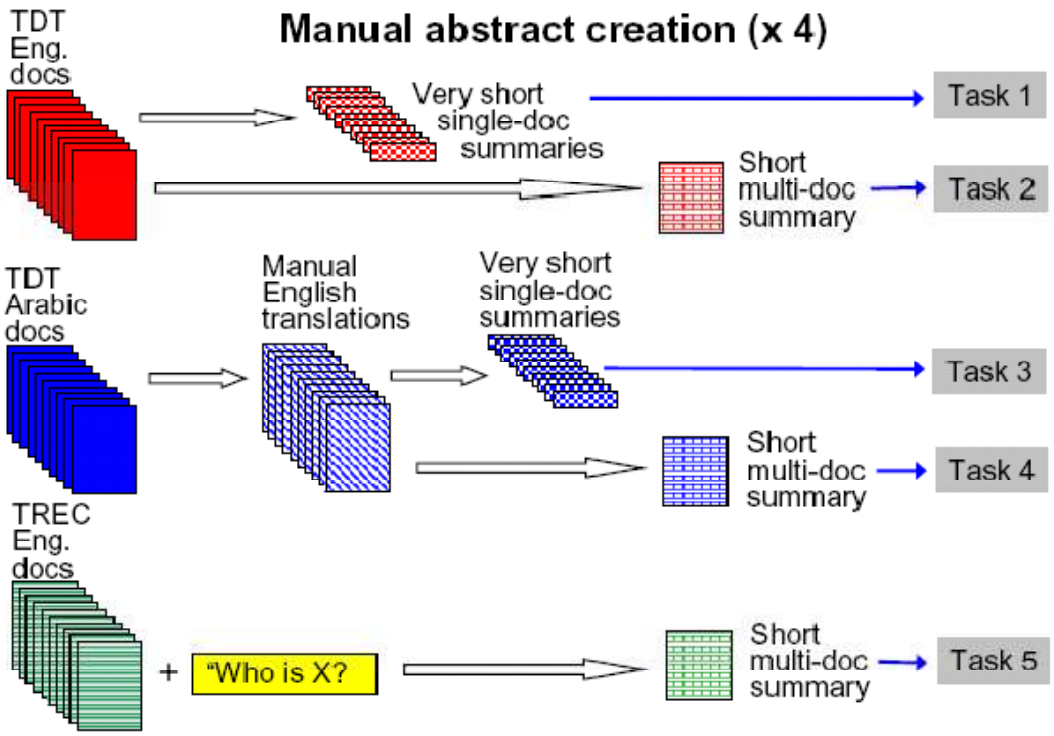
- Task 1
  - VS summary of each document in a cluster
  - Baseline = first 75 bytes of document
  - Evaluation = ROUGE
- Task 2
  - S summary of a document cluster
  - Baseline = first 665 bytes of most recent document
  - Evaluation = ROUGE

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

SLIDE FROM Document Understanding Conferences

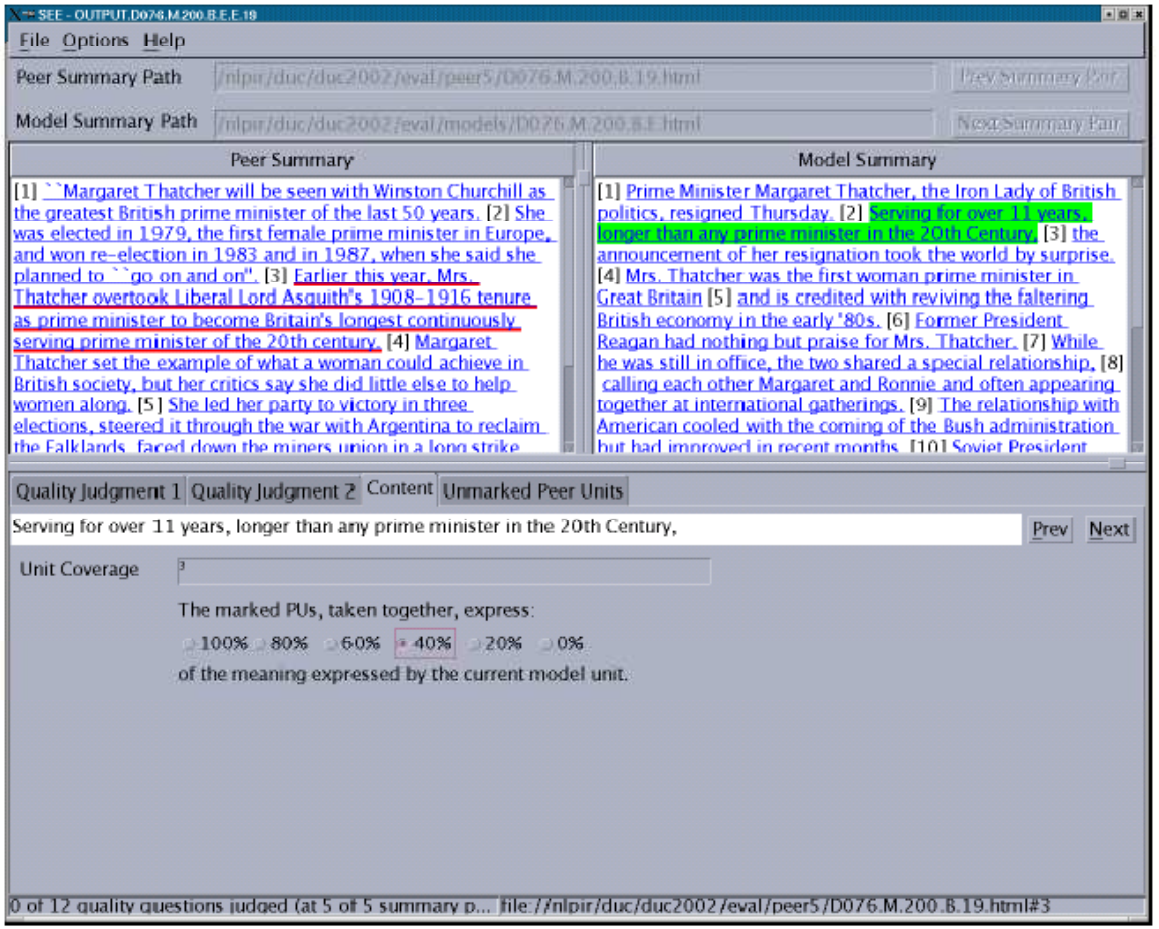
Summary of tasks



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

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)

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

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

ROUGE package

- ROUGE-L: Based on longest common subsequence
- 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

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-2: 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

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

SUMMAC tasks

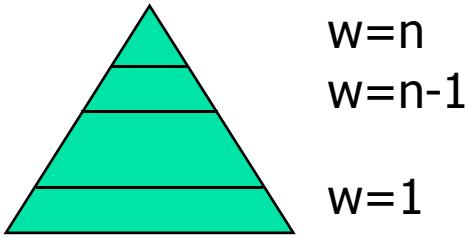
- Categorization task
  - generic summaries
  - given n categories and a summary, the analyst has to classify the document in one of the n categories or none of them
  - one wants to measure whether summaries reduce classification time without loosing classification accuracy

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
    - "13 sailors have been killed" ~ "rebels killed 13 people"
- 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

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

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

$$Score = D / Max$$

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

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

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 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

Summarization with SUMMA

- 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
- Components to 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
- Centroid of sets of documents created
- N-gram computation and N-gram similarity computation

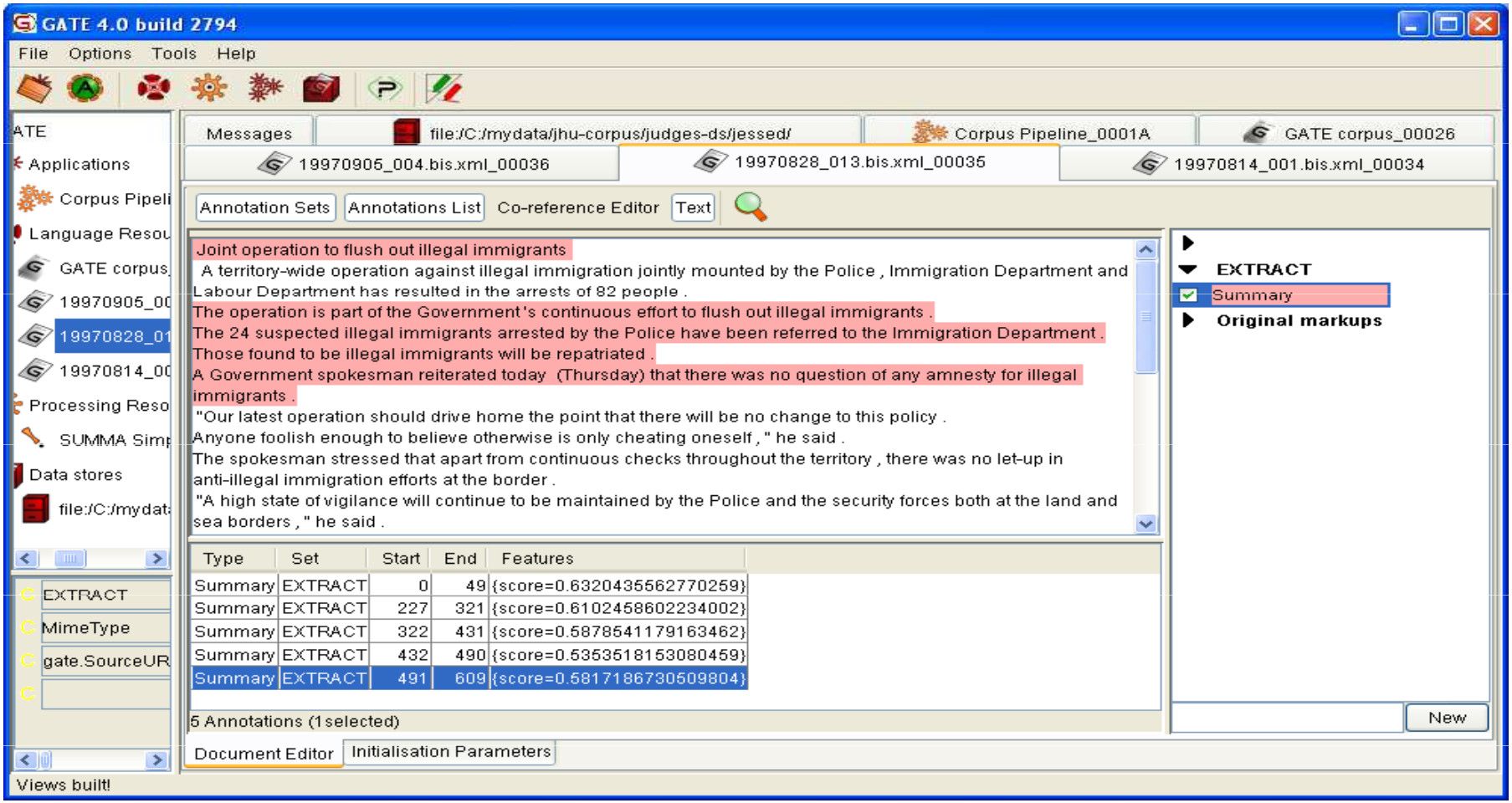
Feature Computation (some)

- 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

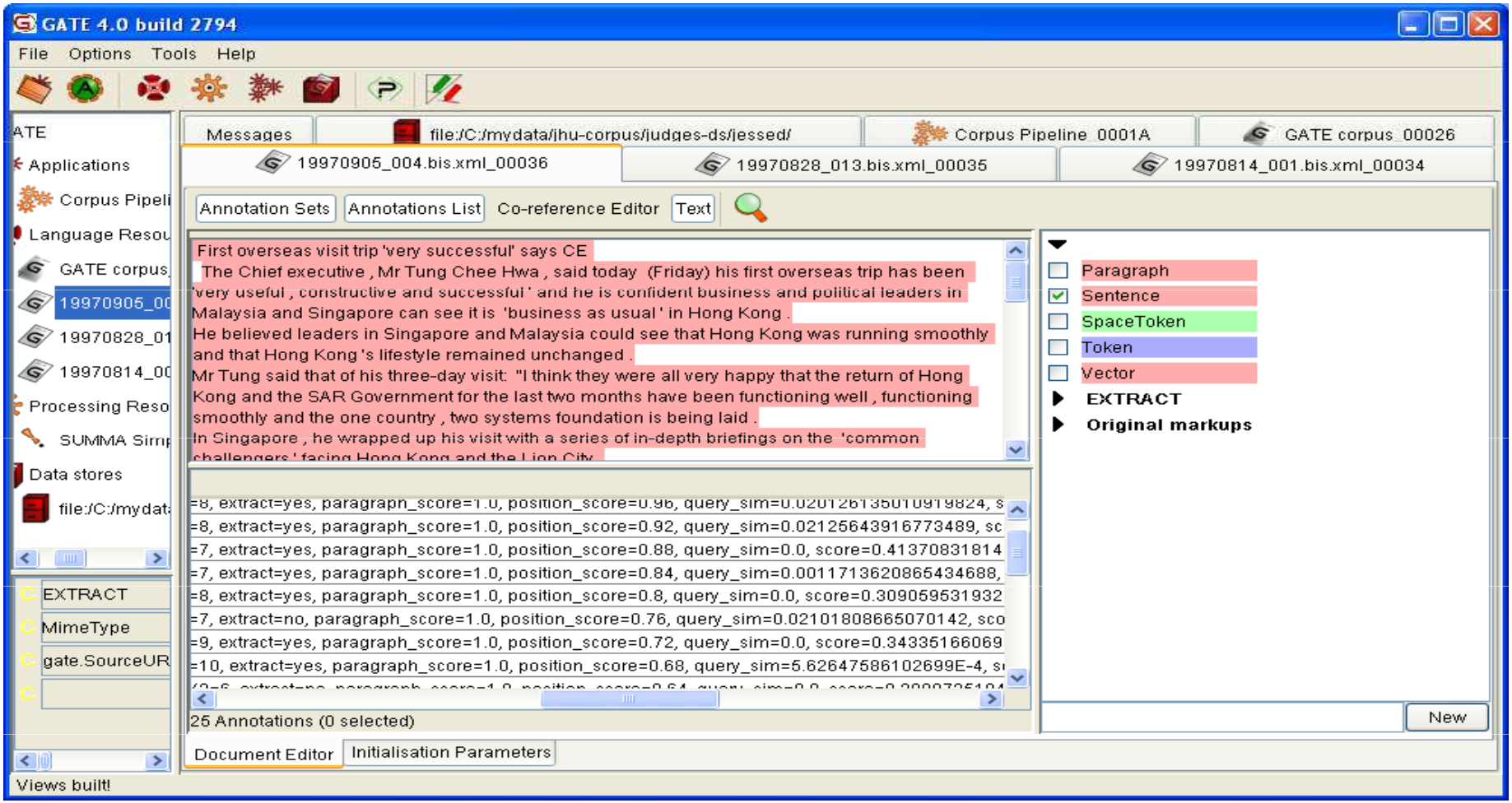
Applications

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

Sentences selected for summary



Features computed for each sentence



Summarizer can be trained

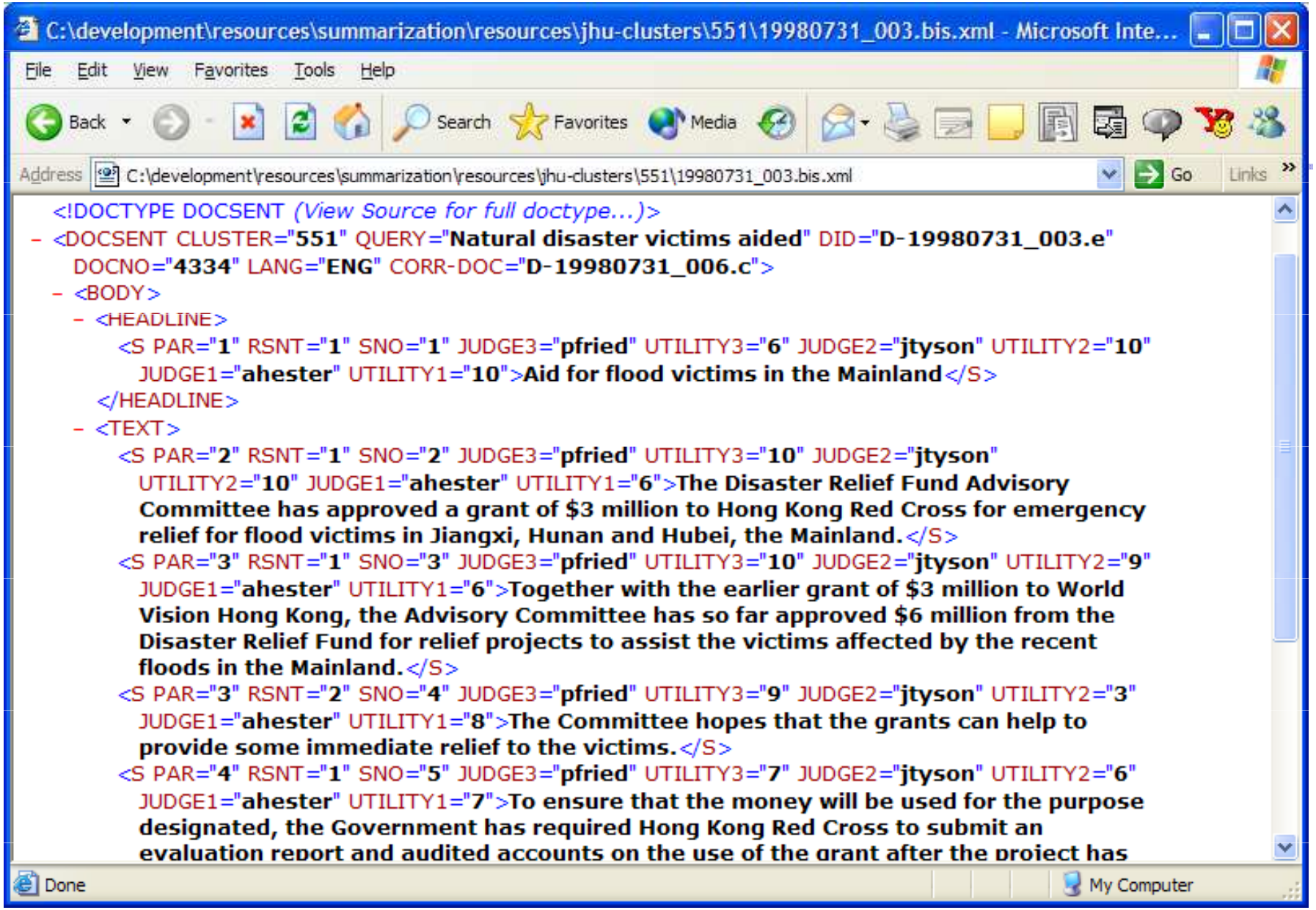
- GATE incorporates ML functionalities through WEKA (Witten&Frank'99) and LibSVM package (http://www.csie.ntu.edu.tw/~cjlin/libsvm)
- training and testing modes are available
  - annotate sentences selected by humans as keys (this can be done with a number of resources to be presented)
  - annotate sentences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

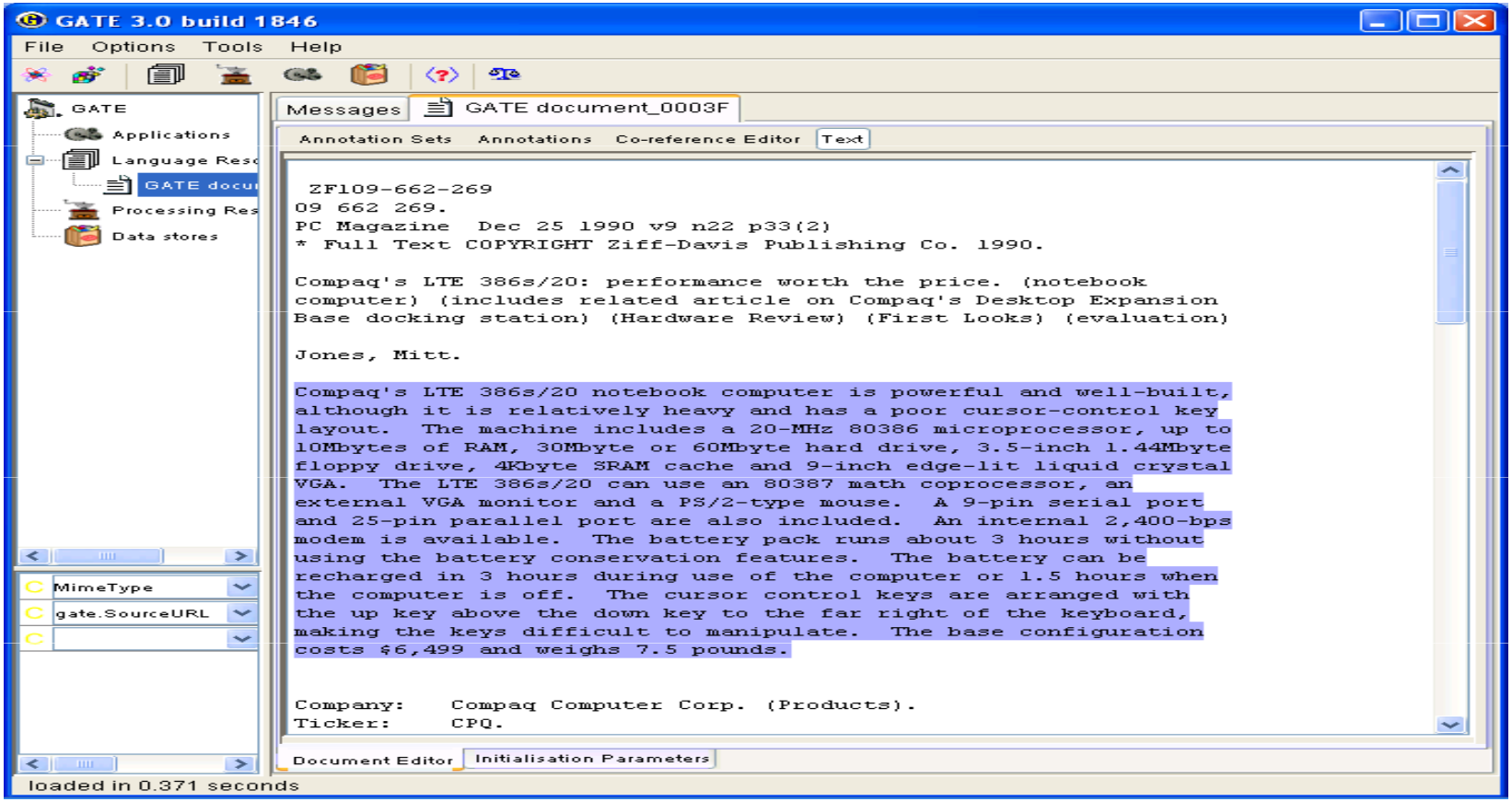
- Hong Kong News Corpus
- formatted in XML
- 40 topics/themes identified by LDC
- creation of a list of relevant documents for each topic
- 10 documents selected for each topic = clusters
- 3 judges evaluate each sentence in each document
- relevance judgements associated to each sentence (relative utility)
- 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.)


C:\development\resources\summarization\resources\jhu-clusters\551\19980731_003.bis.xml - Microsoft Inte...
File Edit View Favorites Tools Help
Back Search Favorites Media
Address C:\development\resources\summarization\resources\jhu-clusters\551\19980731_003.bis.xml
Go Links
<!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
Done
My Computer

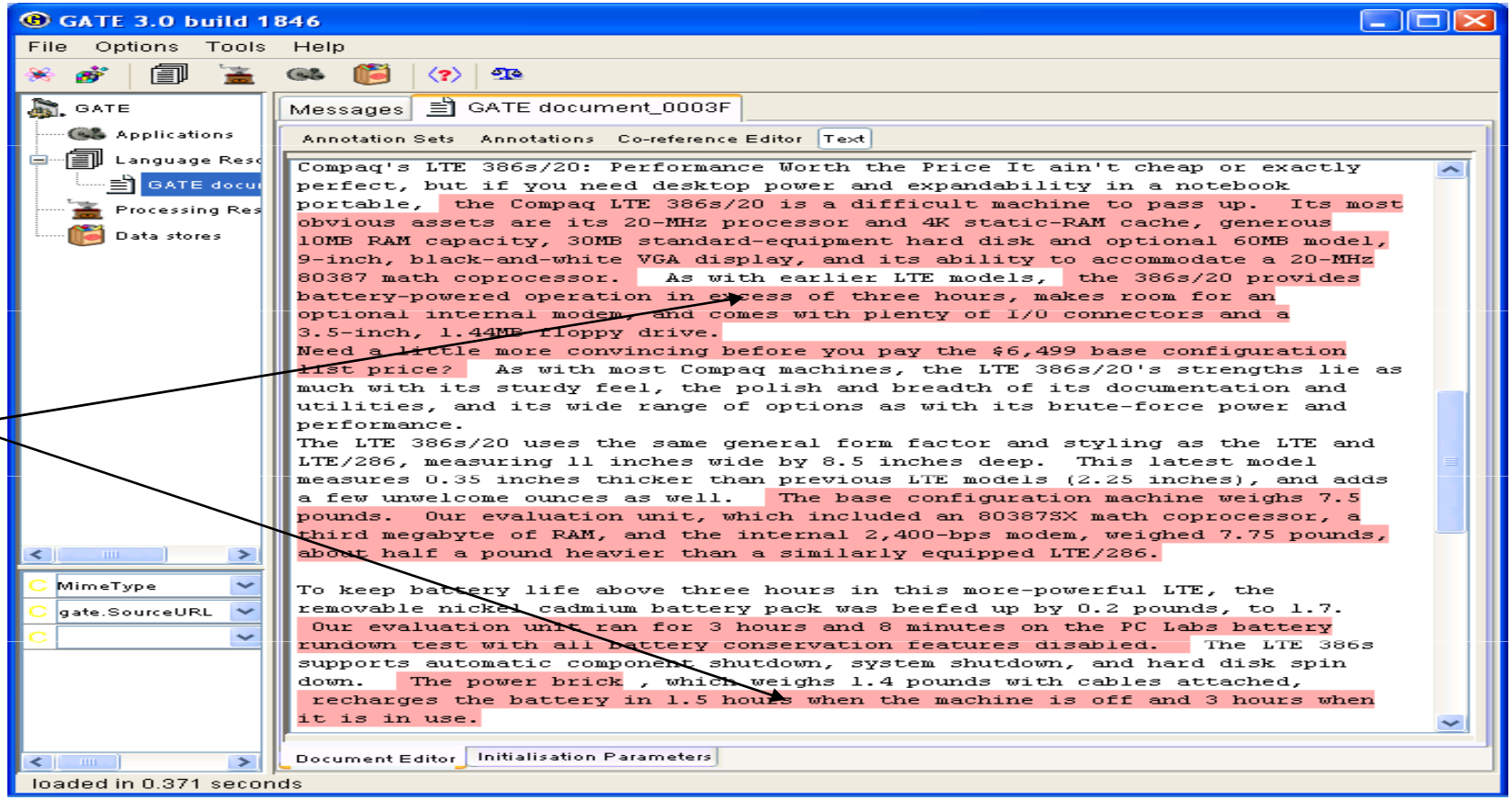
Ziff-Davis Corpus for Summarization

- 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.

Document Summary



Clause Extract



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

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