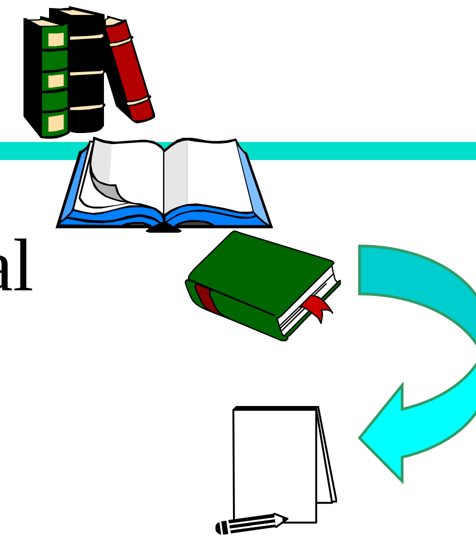
Tutorial:
Automated Text summarization

Eduard Hovy, Chin-Yew Lin, and Daniel Marcu

Information Sciences Institute
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
{hovy,cyl,marcu}@isi.edu
http://www.isi.edu/natural-language/people/{hovy,cyl,marcu}.html

an exciting challenge...



...put a book on the scanner, turn the dial to ‘2 pages’, and read the result...

...download 1000 documents from the web, send them to the summarizer, and select the best ones by reading the summaries of the clusters...

...forward the Japanese email to the summarizer, select ‘1 par’, and skim the translated summary.

Headline news — informing

TV-GUIDES — decision making

2:30am **VC2 – 76**

The Jackal

Movie: Bruce Willis excels as "The Jackal," a cunning assassin who uses many disguises in this 1997 thriller. Richard Gere and Sidney Poitier costar as players from different sides of the law who unite to stop him.

3:00am **KCOP – 13**

The Untouchables

Movie: Eliot Ness (Kevin Costner) and "The Untouchables" take on Robert De Niro's flamboyant Al Capone in the pulse–pounding 1987 adaptation of the popular TV series.\Sean Connery won an Oscar as the Irish beat cop who shows Ness "the Chicago way."\ Brian De Palma directed the feature;\David Mamet wrote the script.\And yes, film majors, the scene at Union Station was lifted directly from the

3:05am **STARZ – 25**

Grosse Pointe Blank

Movie: A razor–sharp script and a fine turn by John Cusack as a troubled hit man mark 1997's "Grosse Pointe Blank," a dark comedy in which the assassin encounters his old flame (Minnie Driver of "Good Will Hunting") at a high–school reunion. Cusack's sister Joan ("In and Out") is hilarious as the killer's devoted assistant, and Alan Arkin makes the most of his small role as Cusack's terrified the

Abstracts of papers — time saving

An Incremental Interpreter for High-Level Programs with Sensing

Giuseppe De Giacomo
Dipartimento di Informatica e Sistemistica
Universita di Roma "La Sapienza"
Via Salaria 113, 00198 Rome, Italy
degiacomo@dis.uniroma1.it

Hector Levesque
Department of Computer Science
University of Toronto
Toronto, Canada M5S 3H5
hector@cs.toronto.edu

Abstract

Like classical planning, the execution of high-level agent programs requires a reasoner to look all the way to a final goal state before even a single action can be taken in the world. This deferral is a serious problem in practice for large programs. Furthermore, the problem is compounded in the presence of sensing actions which provide necessary information, but only after they are executed in the world. To deal with this, we propose (characterize formally in the situation calculus, and implement in Prolog) a new incremental way of interpreting such high-level programs and a new high-level language construct, which together, and without loss of generality, allow much more control to be exercised over when actions can be executed. We argue that such a scheme is the only practical way to deal with large agent programs containing both nondeterminism and sensing.

Introduction

In [4] it was argued that when it comes to providing high level control to autonomous agents or robots, the notion of *high-level program execution* offers an alternative to classical planning that may be more practical in many applications. Briefly, instead of looking for a sequence of actions $\vec{a}$ such that

$$Axioms \models Legal(do(\vec{a}, S_0)) \wedge \phi(do(\vec{a}, S_0))$$

where ϕ is the goal being planned for, we look for a sequence $\vec{a}$ such that

$$Axioms \models Do(\delta, S_0, do(\vec{a}, S_0))$$

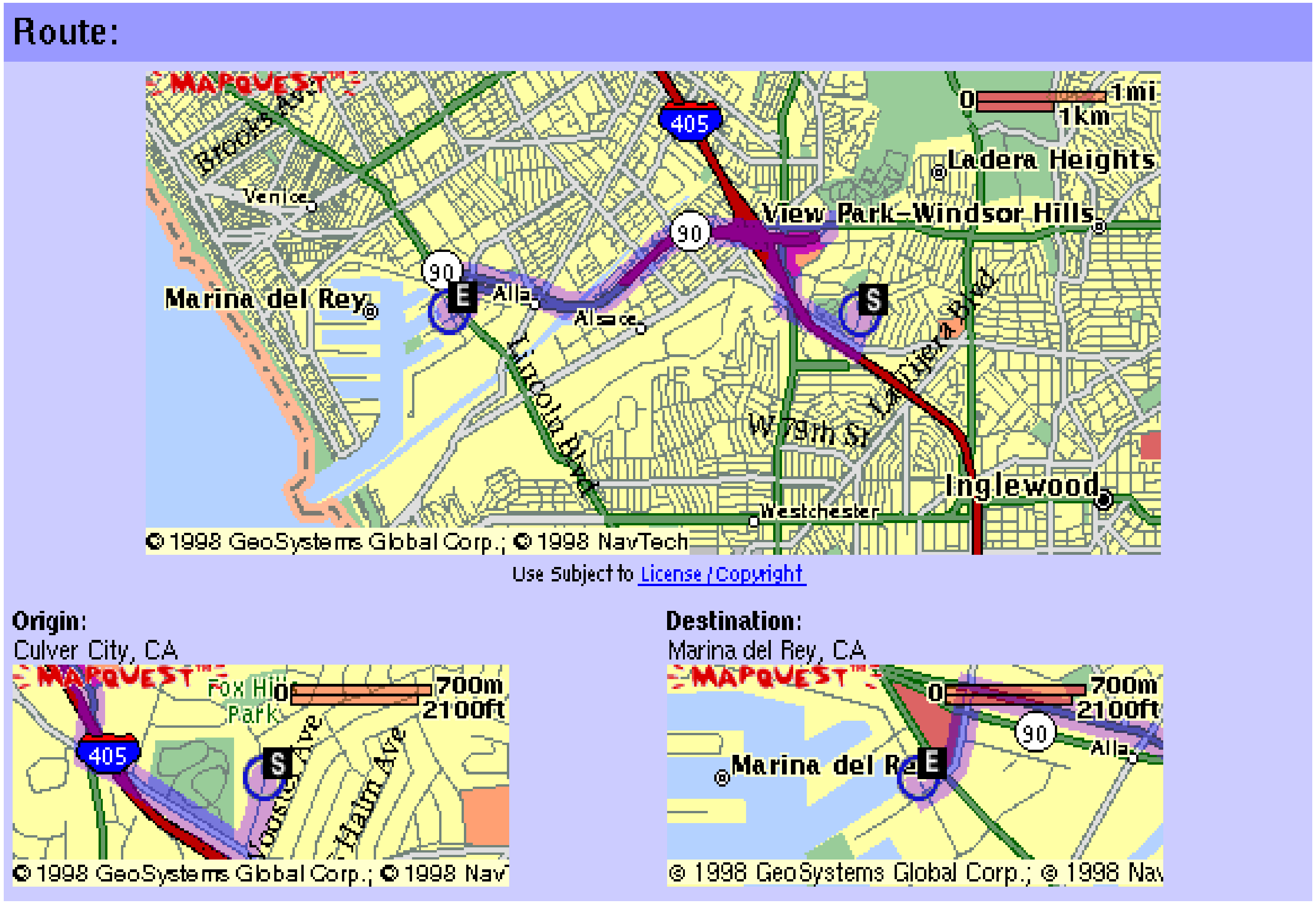
to find a sequence with the right properties. This can involve considerable search when δ is very nondeterministic, but much less search when δ is more deterministic. The feasibility of this approach for AI purposes clearly depends on the expressive power of the programming language in question. In [4], a language called CONGOLOG is presented, which in addition to nondeterminism, contains facilities for sequence, iteration, conditionals, concurrency, and prioritized interrupts. In this paper, we extend the expressive power of this language by providing much finer control over the nondeterminism, and by making provisions for sensing actions. To do so in a way that will be practical even for very large programs requires introducing a different style of on-line program execution.

In the rest of this section, we discuss on-line and off-line execution informally, and show why sensing actions and nondeterminism together can be problematic. In the following section, we formally characterize program execution in the language of the situation calculus. Next, we describe an incremental interpreter in Prolog that is correct with respect to this specification. The final section contains discussion and conclusions.

Off-line and On-line execution

To be compatible with planning, the CONGOLOG interpreter presented in [4] executes in an *off-line* manner, in the sense that it must find a sequence of actions constituting an entire legal execution of a program *before* actually executing any of them in the world.[1] Consider, for example, the following program:

Graphical maps — orienting



Textual Directions — planning

Door to Door Directions:

From:	6420 Green Valley Circle Culver City, CA	
To:	4676 Admiralty Way Marina del Rey, CA	

Direction	**Distance**
1: Start out going South on GREEN VALLEY CIR towards W CENTINELA AVE.	0.2 miles
2: Turn RIGHT onto S CENTINELA AVE.	0.5 miles
3: Turn RIGHT onto SEPULVEDA BLVD.	0.6 miles
4: Turn RIGHT onto W SLAUSON AVE.	0.3 miles
5: Take the CA-90 WEST ramp.	0.1 miles
6: Merge onto CA-90 W.	2.9 miles
7: Turn LEFT onto MINDANAO WAY.	0.3 miles
8: Turn RIGHT onto ADMIRALTY WAY.	0.0 miles

Total Distance:	4.9
Estimated Time:	11 minutes

Cliff notes — Laziness support

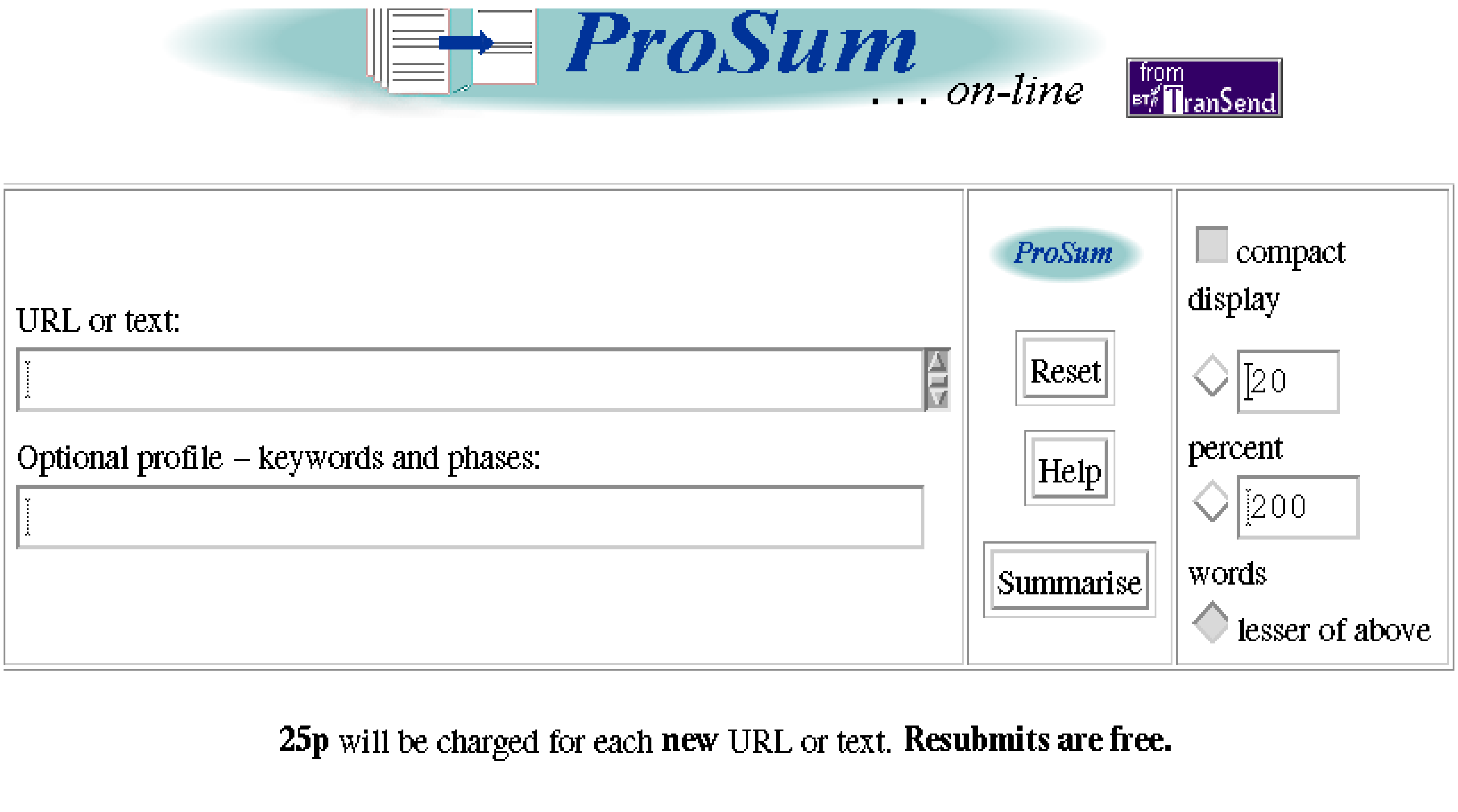
Cliff Notes for the Grapes of Wrath

Posted by Derek on December 02, 1997 at 11:35:43:

In Reply to: Re: I need cliff notes or a summary to TO KILL A MOCKING>> posted by kandice on September 28, 1997 at 20:40:48:

Say can you send me some cliff notes for the grapes of wrath by Wednesday December 3, 1997. I would appricate it very much and I would recomend this page to all my friends so we could ace our english tests on the grapes of wrath. PlEASE SEND ME A COPY OF THE GRAPES OF WRATH CLIFF NOTES I NEED THEM BAD!!!!!!!!!!!!!!

Real systems — Money making

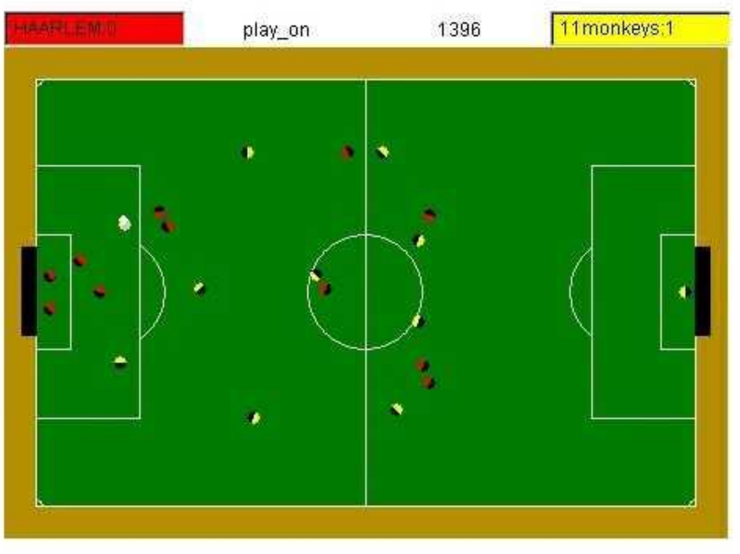


Soccer Game Summaries

HAARLEM offense collapses in stunning defeat at the hands of 11monkeys!

11monkeys displayed their offensive and defensive prowess, shutting out their opponents 7-0. 11monkeys pressed the attack very hard against the HAARLEM defense, keeping the ball in their half of the field for 84% of the game and allowing ample scoring opportunities. HAARLEM pulled their defenders back to stop the onslaught, but to no avail. To that effect, 11monkeys was able to get past HAARLEM's last defender, creating 2 situations where only the goalie was left to defend the net. 11monkeys also handled the ball better, keeping control of the ball for 86% of the game. HAARLEM had a tendency to keep the ball towards the center of the field as well, which may have helped lead them to ruin given the ferocity of the 11monkeys attack.

11monkeys scored using their dribbling technique for 7 of their goals. HAARLEM did not keep a good amount of distance between their players. 11monkeys displayed some of their ball control skills. HAARLEM had their last defender bypassed 2 times for 1 goals.



- AI Agent plans summary
- Winner of prize at IJCAI (Tambe et al., 1999)

Questions

- **What kinds** of summaries do people want?
  - What are *summarizing*, *abstracting*, *gisting*,...?
- **How sophisticated** must summ. systems be?
  - Are statistical techniques sufficient?
  - Or do we need symbolic techniques and deep understanding as well?
- **What milestones** would mark quantum leaps in summarization theory and practice?
  - How do we measure summarization quality?

Table of contents

Examples of Genres

Exercise: summarize the following texts for the following readers:

text1: Coup Attempt	**reader1:** your friend, who knows nothing about South Africa.
	reader2: someone who lives in South Africa and knows the political position.
text2: childrens' story	**reader3:** your 4-year-old niece.
	reader4: the Library of Congress.

90 Soldiers Arrested After Coup Attempt In Tribal Homeland
MMABATHO, South Africa (AP)

About 90 soldiers have been arrested and face possible death sentences stemming from a coup attempt in Bophuthatswana, leaders of the tribal homeland said Friday.

Rebel soldiers staged the takeover bid Wednesday, detaining homeland President Lucas Mangope and several top Cabinet officials for 15 hours before South African soldiers and police rushed to the homeland, rescuing the leaders and restoring them to power.

At least three soldiers and two civilians died in the uprising.

Bophuthatswana's Minister of Justice G. Godfrey Mothibe told a news conference that those arrested have been charged with high treason and if convicted could be sentenced to death. He said the accused were to appear in court Monday.

All those arrested in the coup attempt have been described as young troops, the most senior being a warrant officer.

During the coup rebel soldiers installed as head of state Rocky Malebane-Metsing, leader of the opposition Progressive Peoples Party.

Malebane-Metsing escaped capture and his whereabouts remained unknown, officials said. Several unsubstantiated reports said he fled to nearby Botswana.

Warrant Officer M.T.F. Phiri, described by Mangope as one of the coup leaders, was arrested Friday in Mmabatho, capital of the nominally independent homeland, officials said.

Bophuthatswana, which has a population of 1.7 million spread over seven separate land blocks, is one of 10 tribal homelands in South Africa. About half of South Africa's 26 million blacks live in the homelands, none of which are recognized internationally.

Hennie Riekert, the homeland's defense minister, said South African troops were to remain in Bophuthatswana but will not become a ``permanent presence.''

Bophuthatswana's Foreign Minister Solomon Rathebe defended South Africa's intervention.

``The fact that ... the South African government (was invited) to assist in this drama is not anything new nor peculiar to Bophuthatswana,'' Rathebe said. ``But why South Africa, one might ask? Because she is the only country with whom Bophuthatswana enjoys diplomatic relations and has formal agreements.''

Mangope described the mutual defense treaty between the homeland and South Africa as ``similar to the NATO agreement,'' referring to the Atlantic military alliance. He did not elaborate.

Asked about the causes of the coup, Mangope said, ``We granted people freedom perhaps ... to the extent of planning a thing like this.''

The uprising began around 2 a.m. Wednesday when rebel soldiers took Mangope and his top ministers from their homes to the national sports stadium.

On Wednesday evening, South African soldiers and police stormed the stadium, rescuing Mangope and his Cabinet.

South African President P.W. Botha and three of his Cabinet ministers flew to Mmabatho late Wednesday and met with Mangope, the homeland's only president since it was declared independent in 1977.

The South African government has said, without producing evidence, that the outlawed African National Congress may be linked to the coup.

The ANC, based in Lusaka, Zambia, dismissed the claims and said South Africa's actions showed that it maintains tight control over the homeland governments. The group seeks to topple the Pretoria government.

The African National Congress and other anti-government organizations consider the homelands part of an apartheid system designed to fragment the black majority and deny them political rights in South Africa.

If You Give a Mouse a Cookie

Laura Joffe Numeroff © 1985

If you give a mouse a cookie,he’s going to ask for a glass of milk.
When you give him the milk, he’ll probably ask you for a straw.
When he’s finished, he’ll ask for a napkin.
Then he’ll want to look in the mirror to make sure he doesn’t have a milk mustache.
When he looks into the mirror, he might notice his hair needs a trim.
So he’ll probably ask for a pair of nail scissors.
When he’s finished giving himself a trim, he’ll want a broom to sweep up.
He’ll start sweeping.
He might get carried away and sweep every room in the house.
He may even end up washing the floors as well.
When he’s done, he’ll probably want to take a nap.
You’ll have to fix up a little box for him with a blanket and a pillow.
He’ll crawl in, make himself comfortable, and fluff the pillow a few times.
He’ll probably ask you to read him a story.
When you read to him from one of your picture books, he'll ask to see the pictures.
When he looks at the pictures, he’ll get so excited that he’ll want to draw one of his own. He’ll ask for paper and crayons.
He’ll draw a picture. When the picture is finished, he’ll want to sign his name, with a pen.
Then he’ll want to hang his picture on your refrigerator. Which means he’ll need Scotch tape.
He’ll hang up his drawing and stand back to look at it. Looking at the refrigerator will remind him that he’s thirsty.
So…he’ll ask for a glass of milk.
And chances are that if he asks for a glass of milk, he’s going to want a cookie to go with it.

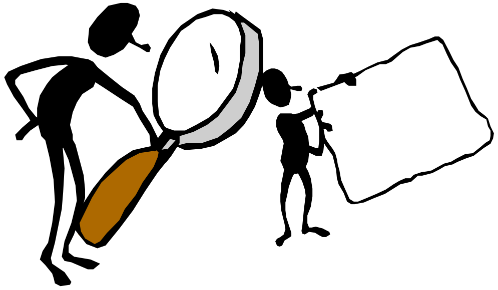
Aspects that Describe Summaries

- **Input** (Sparck Jones 97,Hovy and Lin 99)
  - Single-document *vs*. multi-document...*fuse together texts?*
  - Domain-specific *vs*. general...*use domain-specific techniques?*
  - Genre...*use genre-specific (newspaper, report…) techniques?*
  - Scale and form…*input large or small? Structured or free-form?*
  - Monolingual *vs*. multilingual...*need to cross language barrier?*
- **Purpose**
  - Situation...*embedded in larger system (MT, IR) or not?*
  - Generic *vs*. query-oriented...*author's view or user's interest?*
  - Indicative *vs*. informative...*categorization or understanding?*
  - Background *vs*. just-the-news…*does user have prior knowledge?*
- **Output**
  - Extract *vs*. abstract...*use text fragments or re-phrase content?*
  - Domain-specific *vs*. general...*use domain-specific format?*
  - Style…*make informative, indicative, aggregative, critical...*

Table of contents

Two Psycholinguistic Studies

- Coarse-grained summarization protocols from professional summarizers (Kintsch and van Dijk, 78):
  - Delete material that is trivial or redundant.
  - Use superordinate concepts and actions.
  - Select or invent topic sentence.

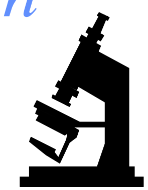


- 552 finely-grained summarization strategies from professional summarizers (Endres-Niggemeyer, 98):
  - **Self control**: make yourself feel comfortable.
  - **Processing**: produce a unit as soon as you have enough data.
  - **Info organization**: use “Discussion” section to check results.
  - **Content selection**: the table of contents is relevant.

Computational Approach: Basics

Top-Down:

- *I know what I want! — don't confuse me with drivel!*



- User wants only certain types of info.
- System needs *particular criteria of interest*, used to focus search.

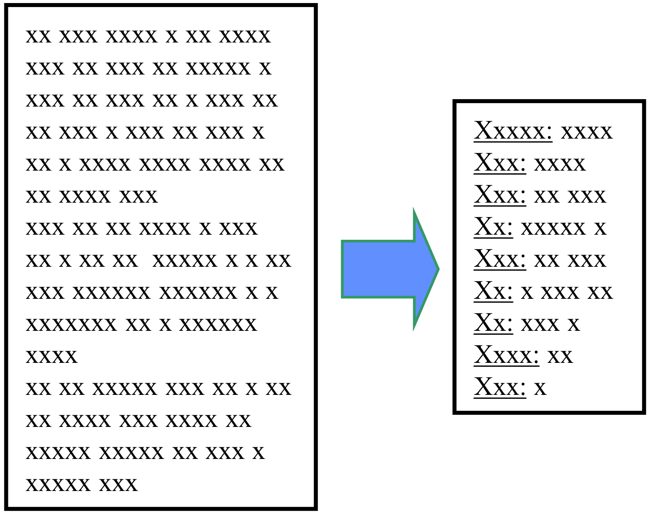
Bottom-Up:

- *I'm dead curious: what's in the text?*

- User wants anything that's important.
- System needs *generic importance metrics*, used to rate content.

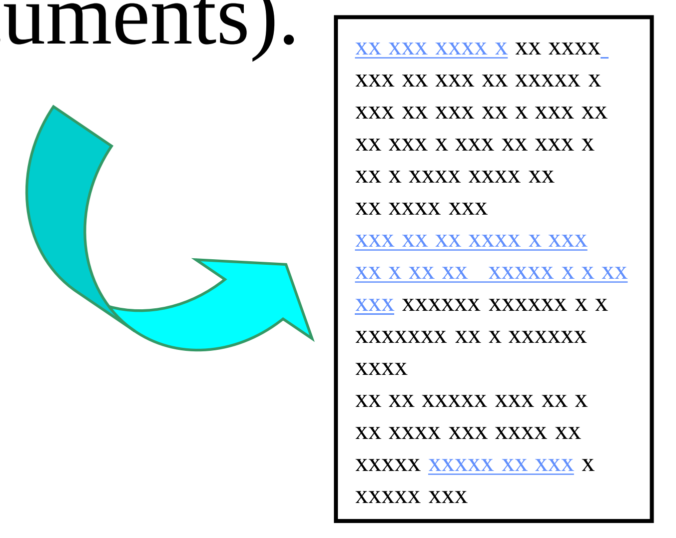
Top-Down: Info. Extraction (IE)

- **IE task**: Given a form and a text, find all the information relevant to each slot of the form and fill it in.
- **Summ-IE task**: Given a query, select the best form, fill it in, and generate the contents.
- **Questions**:
  1. IE works only for very particular forms; can it scale up?
  2. What about info that doesn't fit into any form—is this a generic limitation of IE?



Bottom-Up: Info. Retrieval (IR)

- **<u>IR task</u>**: Given a query, find the relevant document(s) from a large set of documents.
- **<u>Summ-IR task</u>**: Given a query, find the relevant passage(s) from a set of passages (i.e., from one or more documents).
- **<u>Questions</u>**:
  1. IR techniques work on large volumes of data; can they scale down accurately enough?
  2. IR works on words; do abstracts require abstract representations?



Paradigms: IE vs. IR

IE:	IR:
• **Approach**: try to 'understand' text—transform content into 'deeper' notation; then manipulate that.	• **Approach**: operate at word level—use word frequency, collocation counts, etc.
• **Need**: rules for text analysis and manipulation, at all levels.	• **Need**: large amounts of text.
• **Strengths**: higher quality; supports abstracting.	• **Strengths**: robust; good for query-oriented summaries.
• **Weaknesses**: speed; still needs to scale up to robust open-domain summarization.	• **Weaknesses**: lower quality; inability to manipulate information at abstract levels.

Deep and Shallow, Down and Up...

Today:

Increasingly, techniques hybridize: people use word-level counting techniques to fill IE forms' slots, and try to use IE-like discourse and quasi-semantic notions in the IR approach.

Thus:

You can use either deep or shallow paradigms for either top-down or bottom-up approaches!

Toward the Final Answer...

- **Problem**: What if neither IR-like nor IE-like methods work?
  - sometimes counting and forms are insufficient,
  - and then you need to do inference to *understand*.

Word counting

Mrs. Coolidge: "What did the preacher preach about?"
Coolidge: "Sin."
Mrs. Coolidge: "What did he say?"
Coolidge: "He's against it."

Inference

- **Solution**:
  - semantic analysis of the text (NLP),
  - using adequate knowledge bases that support inference (AI).

The Optimal Solution…

Combine strengths of both paradigms…

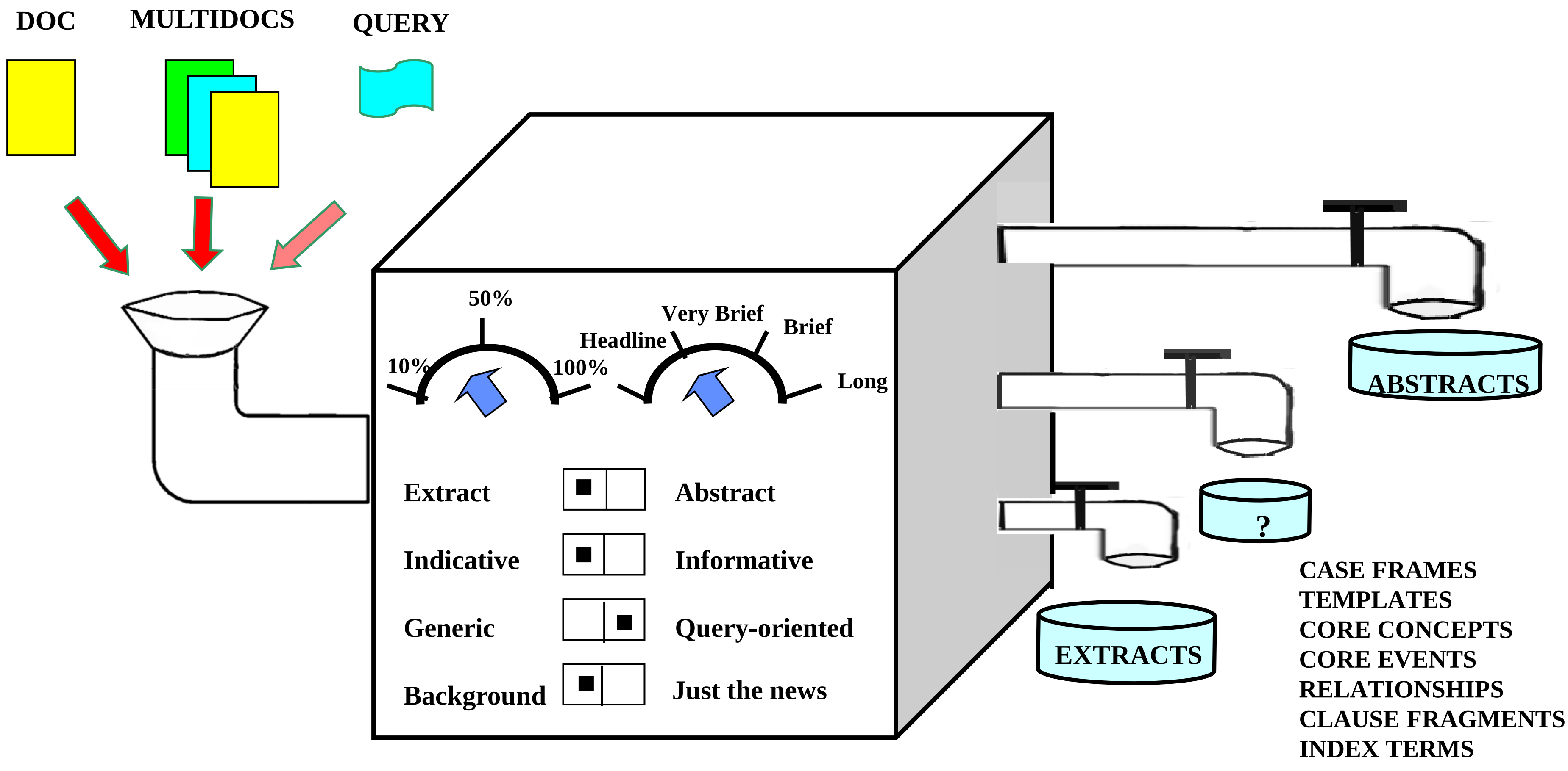
…use IE/NLP when you have suitable form(s),

…use IR when you don't…

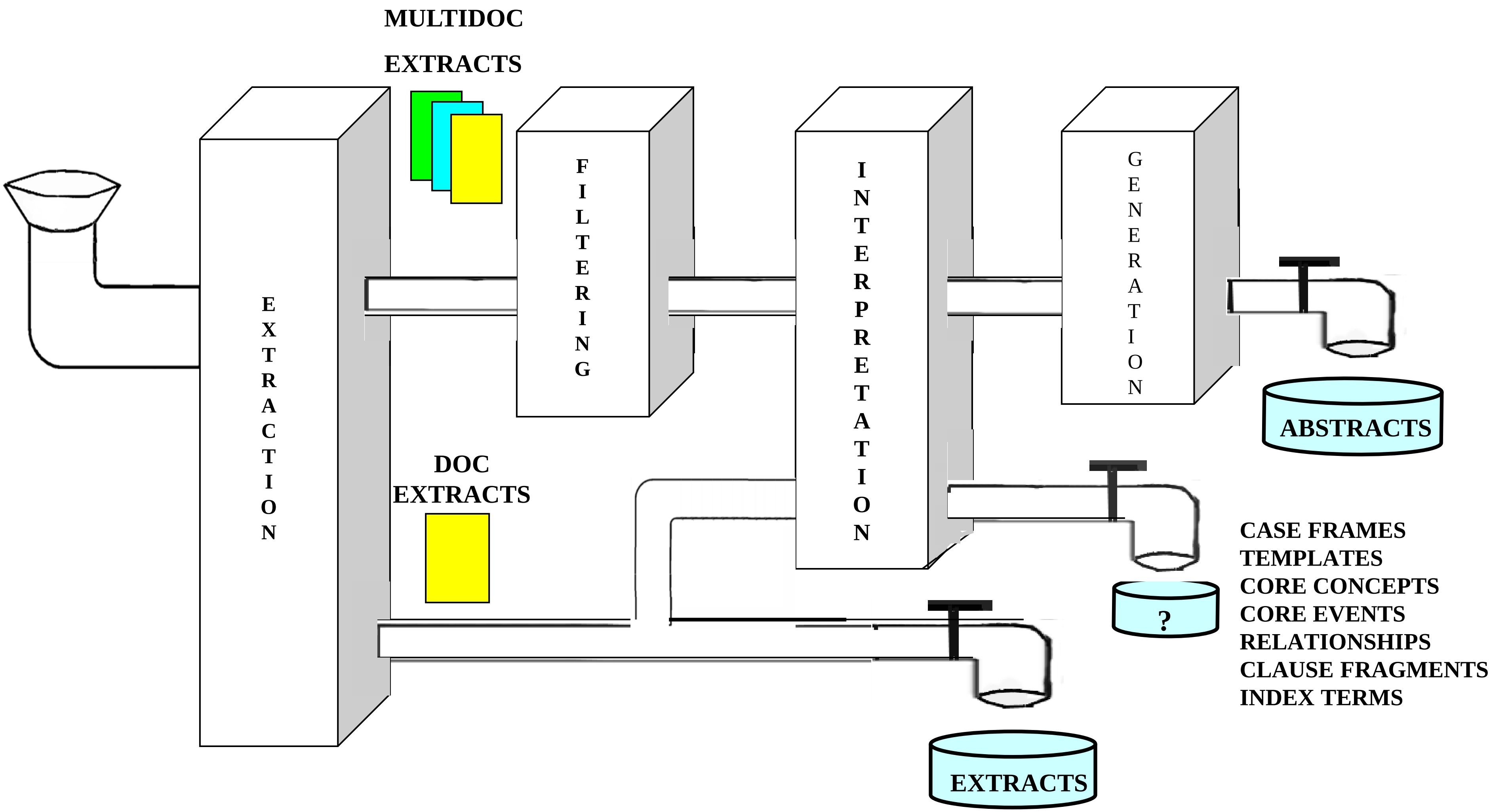
…but how exactly to do it?

Table of contents

A Summarization Machine



The Modules of the Summarization Machine



Typical 3 Stages of Summarization

1. Topic Identification: find/extract the most important material

2. Topic Interpretation: compress it

3. Summary Generation: say it in your own words

...as easy as that!

Some Definitions

- **Language**:
  - **Syntax** = grammar, sentence structure *sleep*
    colorless furiously ideas green — no syntax
  - **Semantics** = meaning
    colorless green ideas sleep furiously — no semantics
- **Evaluation**:
  - **Recall** = how many of the things you should have found/did, did you actually find/do?
  - **Precision** = of those you actually found/did, how many were correct?

Table of contents

1. Motivation.
2. Genres and types of summaries.
3. Approaches and paradigms.
4. Summarization methods (& exercise).
    - Topic Extraction.
    - Interpretation.
    - Generation.
5. Evaluating summaries.
6. The future.

Overview of Extraction Methods

- **General method**: score each sentence; combine scores; choose best sentence(s)
- **Scoring techniques**:
  - Position in the text: lead method; optimal position policy; title/heading method
  - Cue phrases in sentences
  - Word frequencies throughout the text
  - Cohesion: links among words; word co-occurrence; coreference; lexical chains
  - Discourse structure of the text
  - Information Extraction: parsing and analysis

Note

- The recall and precision figures reported here reflect the ability of various methods to match human performance on the task of identifying the sentences/clauses that are important in texts.
- Rely on evaluations using six corpora:

  (Edmundson, 68; Kupiec et al., 95; Teufel and Moens, 97; Marcu, 97; Jing et al., 98; SUMMAC, 98).

Position-based method (1)

- **Claim**: Important sentences occur at the beginning (and/or end) of texts.
- **Lead method**: just take first sentence(s)!
- Experiments:
  - In 85% of 200 individual paragraphs the topic sentences occurred in initial position and in 7% in final position (Baxendale, 58).
  - Only 13% of the paragraphs of contemporary writers start with topic sentences (Donlan, 80).

Position-Based Method (2)

Individual contribution

- (Edmundson, 68)
  - 52% recall & precision in combination with title (25% lead baseline)
- (Kupiec et al., 95)
  - 33% recall & precision
  - (24% lead baseline)
- (Teufel and Moens, 97)
  - 32% recall and precision (28% lead baseline)

Cumulative contribution

- (Edmundson, 68)
  - the best individual method
- (Kupiec et al., 95)
  - the best individual method
- (Teufel and Moens, 97)
  - increased performance by 10% when combined with the cue-based method

Optimum Position Policy (OPP)

Claim: Important sentences are located at positions that are genre-dependent; these positions can be determined automatically through training:

- **Corpus**: 13000 newspaper articles (ZIFF corpus).
- **Step 1**: For each article, enumerate sentence positions (both $\rightarrow$ and $\leftarrow$).
- **Step 2**: For each sentence, determine *yield* (= overlap between sentences and the index terms for the article).
- **Step 3**: Create partial ordering over the locations where sentences containing important words occur: Optimal Position Policy (OPP). (Lin and Hovy, 97)

Opp (cont.)

– OPP for ZIFF corpus:

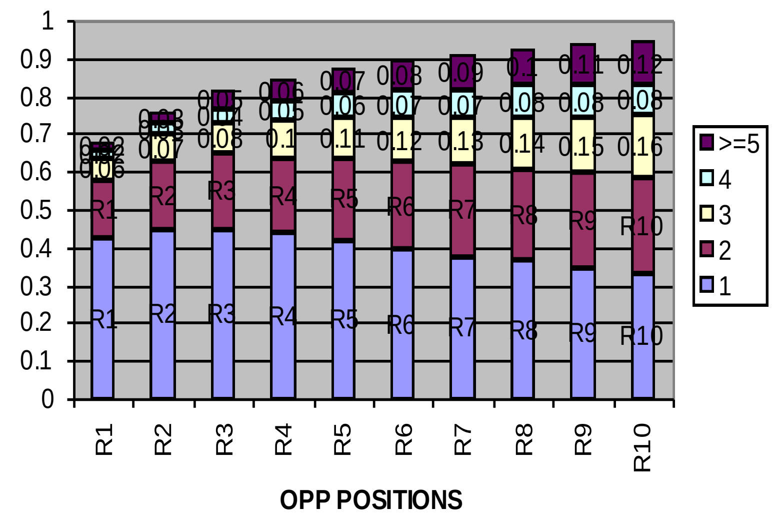
(T) > (P_2,S_1) > (P_3,S_1) > (P_2,S_2) > $\{(P_4,S_1),(P_5,S_1),(P_3,S_2)\}$ >…

(T=title; P=paragraph; S=sentence)

– OPP for *Wall Street Journal*: (T)>(P_1,S_1)>...

– **Results**: testing corpus of 2900 articles: Recall=35%, Precision=38%.

– **Results**: 10%-extracts cover 91% of the salient words.



Position: Title-Based Method (1)

- **Claim**: Words in titles and headings are positively relevant to summarization.

- Shown to be statistically valid at 99% level of significance (Edmundson, 68).
- Empirically shown to be useful in summarization systems.

title-Based Method (2)

Individual contribution

- (Edmundson, 68)
  - 40% recall & precision (25% lead baseline)
- (Teufel and Moens, 97)
  - 21.7% recall & precision (28% lead baseline)

Cumulative contribution

- (Edmundson, 68)
  - increased performance by 8% when combined with the title- and cue-based methods.
- (Teufel and Moens, 97)
  - increased performance by 3% when combined with cue-, location-, position-, and word-frequency-based methods.

Cue-Phrase method (1)

- **Claim 1**: Important sentences contain 'bonus phrases', such as *significantly, In this paper we show*, and *In conclusion*, while non-important sentences contain 'stigma phrases' such as *hardly* and *impossible*.
- **Claim 2**: These phrases can be detected automatically (Kupiec et al. 95; Teufel and Moens 97).
- **Method**: Add to sentence score if it contains a bonus phrase, penalize if it contains a stigma phrase.

Cue-Phrase Method (2)

Individual contribution

- (Edmundson, 68)
  - 45% recall & precision (25% lead baseline)
- (Kupiec et al., 95)
  - 29% recall & precision (24% lead baseline)
- (Teufel and Moens, 97)
  - 55% recall & precision (28% lead baseline)

Cumulative contribution

- (Edmundson, 68)
  - increased performance by 7% when combined with the title and position methods.
- (Kupiec et al., 95)
  - increased performance by 9% when combined with the position method.
- (Teufel and Moens, 97)
  - the best individual method.

Learning Cue Phrases for SUMMARIST

- **Corpora**: newspaper and CompLing articles
- **Several methods**: measure frequencies of words in high-yield sentences in various ways
- **Results**: single and multi-word phrases

Method 1		**Method 2**	
S1	*phrase*	*S2*	*phrase*
7.666	this paper present	3.432	in this paper
7.666	machine learn algorithm	2.889	this paper we
6.909	present the result	2.266	section conclusion
6.888	paper we have	2.279	a set of
6.340	this paper we	2.044	the result of

w_s = *score w in Sum*

w_t = *score w in Text*

df = # texts with w

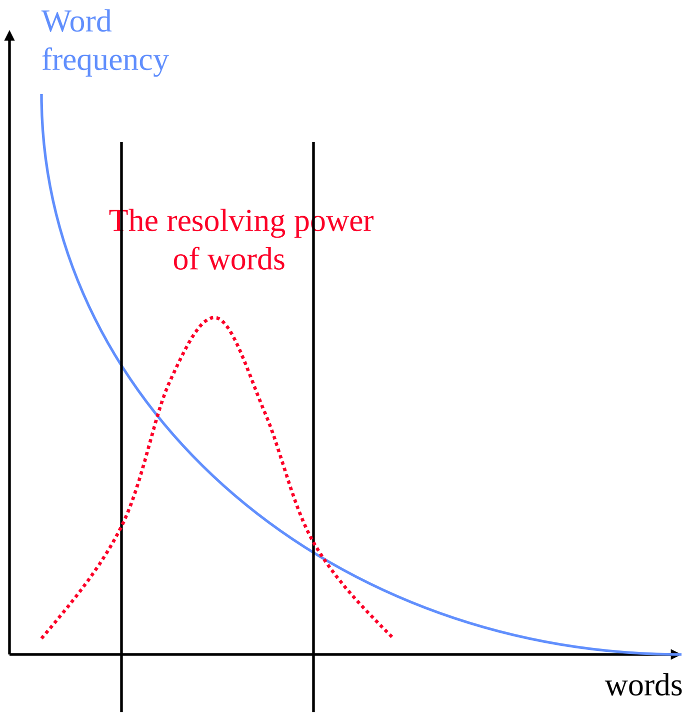
D = total # texts

$S1 = w_s / w_t$

$S2 = w_s / w_t * df / D$

(Liu and Hovy, 98)

Word-frequency-based method (1)



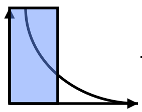
(Luhn, 59)

- **Claim**: Important sentences contain words that occur "somewhat" frequently.
- **Method**: Increase sentence score for each frequent word.
- **Evaluation**: Straightforward approach empirically shown to be mostly detrimental in summarization systems.

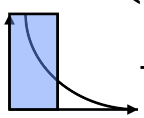
Word-Frequency-Based Method (2)

Individual contribution

- (Edmundson, 68)



  - 36% recall & precision (25% lead baseline)
- (Kupiec et al., 95)



  - 20% recall & precision (24% lead baseline)
- (Teufel and Moens, 97)

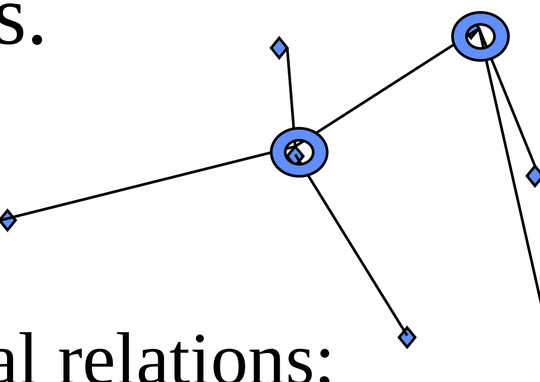
TF-IDF
  - 17% recall & precision (28% lead baseline)

Cumulative contribution

- (Edmundson, 68)
  - decreased performance by 7% when combined with other methods
- (Kupiec et al., 95)
  - decreased performance by 2% when combined...
- (Teufel and Moens, 97)
  - increased performance by 0.2% when combined...

Cohesion-based methods

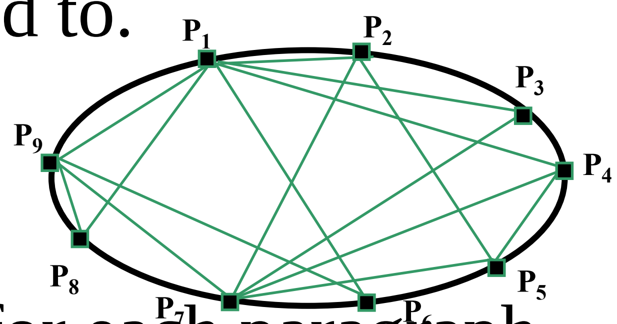
- **Claim**: Important sentences/paragraphs are the highest connected entities in more or less elaborate semantic structures.



- Classes of approaches
  - – word co-occurrences;
  - – local salience and grammatical relations;
  - – co-reference;
  - – lexical similarity (WordNet, lexical chains);
  - – combinations of the above.

Cohesion: WORD co-occurrence (1)

- Apply IR methods at the document level: texts are collections of paragraphs (Salton et al., 94; Mitra et al., 97; Buckley and Cardie, 97):
  - Use a traditional, IR-based, word similarity measure to determine for each paragraph P_i the set S_i of paragraphs that P_i is related to.



- **Method**:
  - determine relatedness score S_i for each paragraph,
  - extract paragraphs with largest S_i scores.

Word co-occurrence method (2)

Study (Mitra et al., 97):

- Corpus: 50 articles from Funk and Wagner Encyclopedia.
- Result: 46.0% overlap between two manual extracts.

	IR-based algorithm	Lead-based algorithm
Optimistic (best overlap)	45.6%	47.9%
Pessimistic (worst overlap)	30.7%	29.5%
Intersection	47.33%	50.0%
Union	55.16%	55.97%

Word co-occurrence method (3)

In the context of query-based summarization

- **Cornell: Smart system approach**
  - expand original query
  - compare expanded query against paragraphs
  - select top three paragraphs (max 25% of original) that are most similar to the original query

  (SUMMAC,98): 71.9% F-score for relevance judgment
- **CGI/CMU approach**
  - maximize query-relevance while minimizing redundancy with previous information (*Maximal Marginal Relevance*)

  (SUMMAC,98): 73.4% F-score for relevance judgment

Cohesion: Local salience Method

- Assumes that important phrasal expressions are given by a combination of grammatical, syntactic, and contextual parameters (Boguraev and Kennedy, 97):

  CNTX: 50 iff the expression is in the current discourse segment
  SUBJ: 80 iff the expression is a subject
  EXST: 70 iff the expression is an existential construction
  ACC: 50 iff the expression is a direct object
  HEAD: 80 iff the expression is not contained in another phrase
  ARG: 50 iff the expression is not contained in an adjunct

- No evaluation of the method.

Cohesion: Lexical chains method (1)

Based on (Morris and Hirst, 91)

But Mr. Kenny's move speeded up work on a **machine** which uses **micro-computers** to control the rate at which an *anaesthetic* is pumped into the blood of *patients* undergoing *surgery*. Such **machines** are nothing new. But Mr. Kenny's **device** uses two **personal-computers** to achieve much closer monitoring of the **pump** feeding the *anaesthetic* into the *patient*. Extensive testing of the **equipment** has sufficiently impressed the authorities which regulate *medical* **equipment** in Britain, and, so far, four other countries, to make this the first such **machine** to be licensed for commercial sale to *hospitals*.

Lexical chains-based method (2)

- Assumes that important sentences are those that are 'traversed' by *strong* chains (Barzilay and Elhadad, 97).
  - Strength(C) = length(C) - #DistinctOccurrences(C)
  - For each chain, choose the first sentence that is traversed by the chain and that uses a representative set of concepts from that chain.

[Jing et al., 98] corpus	LC algorithm		Lead-based algorithm	
	Recall	Prec	Recall	Prec
10% cutoff	67%	61%	82.9%	63.4%
20% cutoff	64%	47%	70.9%	46.9%

Cohesion: Coreference method

- Build co-reference chains (noun/event identity, part-whole relations) between
  - *query and document* - In the context of query-based summarization
  - title and document
  - sentences within document
- Important sentences are those traversed by a large number of chains:
  - a preference is imposed on chains (*query* > title > doc)
- Evaluation: 67% F-score for relevance (SUMMAC, 98). (Baldwin and Morton, 98)

Cohesion: Connectedness method (1)

(Mani and Bloedorn, 97)

- Map texts into graphs:
  - The nodes of the graph are the words of the text.
  - Arcs represent adjacency, grammatical, co-reference, and lexical similarity-based relations.
- Associate importance scores to words (and sentences) by applying the *tf.idf* metric.
- Assume that important words/sentences are those with the highest scores.

Connectedness method (2)

In the context of query-based summarization

- When a query is given, by applying a spreading-activation algorithms, weights can be adjusted; as a results, one can obtain query-sensitive summaries.
- Evaluation (Mani and Bloedorn, 97):
  - IR categorization task: close to full-document categorization results.

[Marcu,97] corpus	TF-IDF method	Spreading activation
10% cutoff F-score	25.2%	32.4%
20% cutoff F-score	35.8%	45.4%

Discourse-based method

- **Claim**: The multi-sentence coherence structure of a text can be constructed, and the 'centrality' of the textual units in this structure reflects their importance.
- Tree-like representation of texts in the style of *Rhetorical Structure Theory* (Mann and Thompson,88).
- Use the discourse representation in order to determine the most important textual units. Attempts:
  - (Ono et al., 1994) for Japanese.
  - (Marcu, 1997,2000) for English.

Rhetorical parsing (Marcu,97)

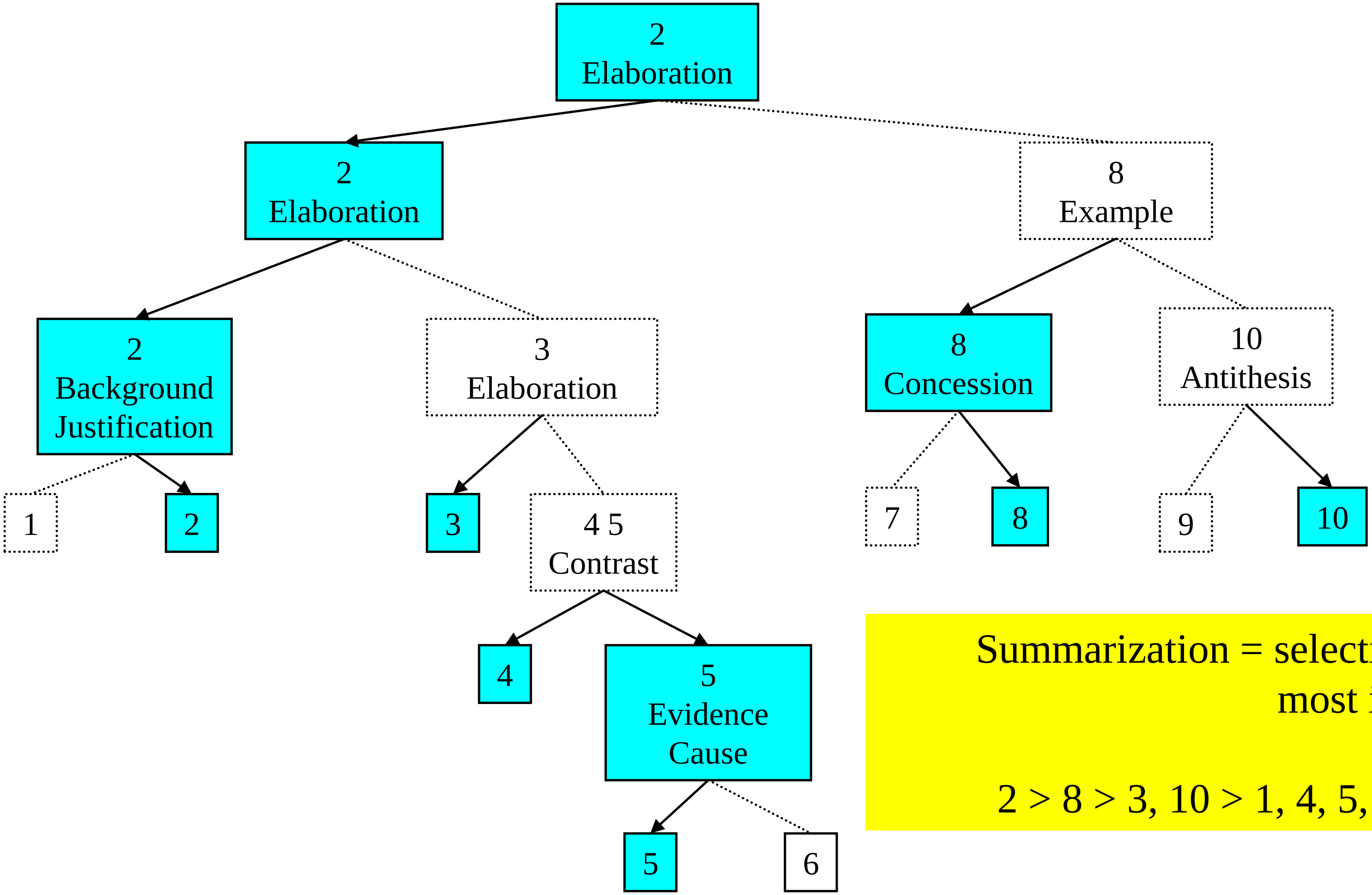
[*With* its distant orbit {– 50 percent farther from the sun than Earth –} and slim atmospheric blanket,[1]] [Mars experiences frigid weather conditions.[2]] [Surface temperatures typically average about –60 degrees Celsius (–76 degrees Fahrenheit) at the equator and can dip to –123 degrees C near the poles.[3]] [Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion,[4]] [*but* any liquid water formed that way would evaporate almost instantly[5]] [*because* of the low atmospheric pressure.[6]]

[*Although* the atmosphere holds a small amount of water, and water-ice clouds sometimes develop,[7]] [most Martian weather involves blowing dust or carbon dioxide.[8]] [Each winter, *for example*, a blizzard of frozen carbon dioxide rages over one pole, and a few meters of this dry-ice snow accumulate as previously frozen carbon dioxide evaporates from the opposite polar cap.[9]] [*Yet* even on the summer pole, {*where* the sun remains in the sky all day long,} temperatures never warm enough to melt frozen water.[10]]

Rhetorical parsing (2)

- Use discourse markers to hypothesize rhetorical relations
  - rhet_rel(CONTRAST, 4, 5) ⊕ rhet_rel(CONTRAT, 4, 6)
  - rhet_rel(EXAMPLE, 9, [7,8]) ⊕ rhet_rel(EXAMPLE, 10, [7,8])
- Use semantic similarity to hypothesize rhetorical relations
  - if similar(u_1,u_2) then
    rhet_rel(ELABORATION, u_2, u_1) ⊕ rhet_rel(BACKGROUND, u_1,u_2)
    else
    rhet_rel(JOIN, u_1, u_2)
  - rhet_rel(JOIN, 3, [1,2]) ⊕ rhet_rel(ELABORATION, [4,6], [1,2])
- Use the hypotheses in order to derive a valid discourse representation of the original text.

Rhetorical parsing (3)



Summarization = selection of the most important units

2 > 8 > 3, 10 > 1, 4, 5, 7, 9 > 6

Discourse method: Evaluation

(using a combination of heuristics for rhetorical parsing disambiguation)

Reduction	Method	Recall	Precision	F-score
10%	Humans	83.20%	75.95%	79.41%
	Program	63.75%	72.50%	67.84%
	Lead	82.91%	63.45%	71.89%
20%	Humans	82.83%	64.93%	72.80%
	Program	61.79%	60.83%	61.31%
	Lead	70.91%	46.96%	56.50%

TREC Corpus
(fourfold cross-validation)

Level	Method	Rec.	Prec.	F-score
Clause	Humans	72.66%	69.63%	71.27%
	Program (training)	67.57%	73.53%	70.42%
	Program (no training)	51.35%	63.33%	56.71%
	Lead	39.68%	39.68%	39.68%
Sentence	Humans	78.11%	79.37%	78.73%
	Program (training)	69.23%	64.29%	66.67%
	Program (no training)	57.69%	51.72%	54.54%
	Lead	54.22%	54.22%	54.22%

Scientific American Corpus

Information Extraction Method (1)

- **Idea**: content selection using forms (templates)
  - Predefine a form, whose slots specify what is of interest.
  - Use a canonical IE system to extract from a (set of) document(s) the relevant information; fill the form.
  - Generate the content of the form as the summary.
- **Previous IE work**:
  - FRUMP (DeJong, 78): '*sketchy scripts*' of terrorism, natural disasters, political visits...
  - (Mauldin, 91): forms for conceptual IR.
  - (Rau and Jacobs, 91): forms for business.
  - (McKeown and Radev, 98): forms for news.

Information Extraction method (2)

- Example form:

MESSAGE:ID	TSL-COL-0001
SECSOURCE:SOURCE	Reuters
SECSOURCE:DATE	26 Feb 93
	Early afternoon
INCIDENT:DATE	26 Feb 93
INCIDENT:LOCATION	World Trade Center
INCIDENT:TYPE	Bombing
HUM TGT:NUMBER	AT LEAST 5

IE State of the Art

- **MUC conferences** (1988–97):
  - Test IE systems on series of domains: Navy sub-language (89), terrorism (92), business (96),...
  - Create increasingly complex form.
  - Evaluate systems, using two measures:
    - *Recall* (how many slots did the system actually fill, out of the total number it should have filled?).
    - *Precision* (how correct were the slots that it filled?).

	1989	1992	1996
Recall	63.9	71.5	67.1
Precision	87.4	84.2	78.3

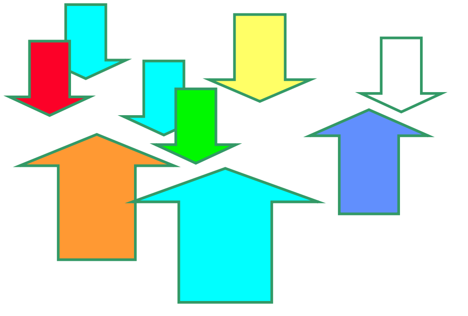
Review of Methods

Bottom-up methods

- Text location: title, position
- Cue phrases.
- Word frequencies
- Internal text cohesion:
  - word co-occurrences
  - local salience
  - co-reference of names, objects
  - lexical similarity
  - semantic rep/graph centrality
- Discourse structure centrality

Top-down methods

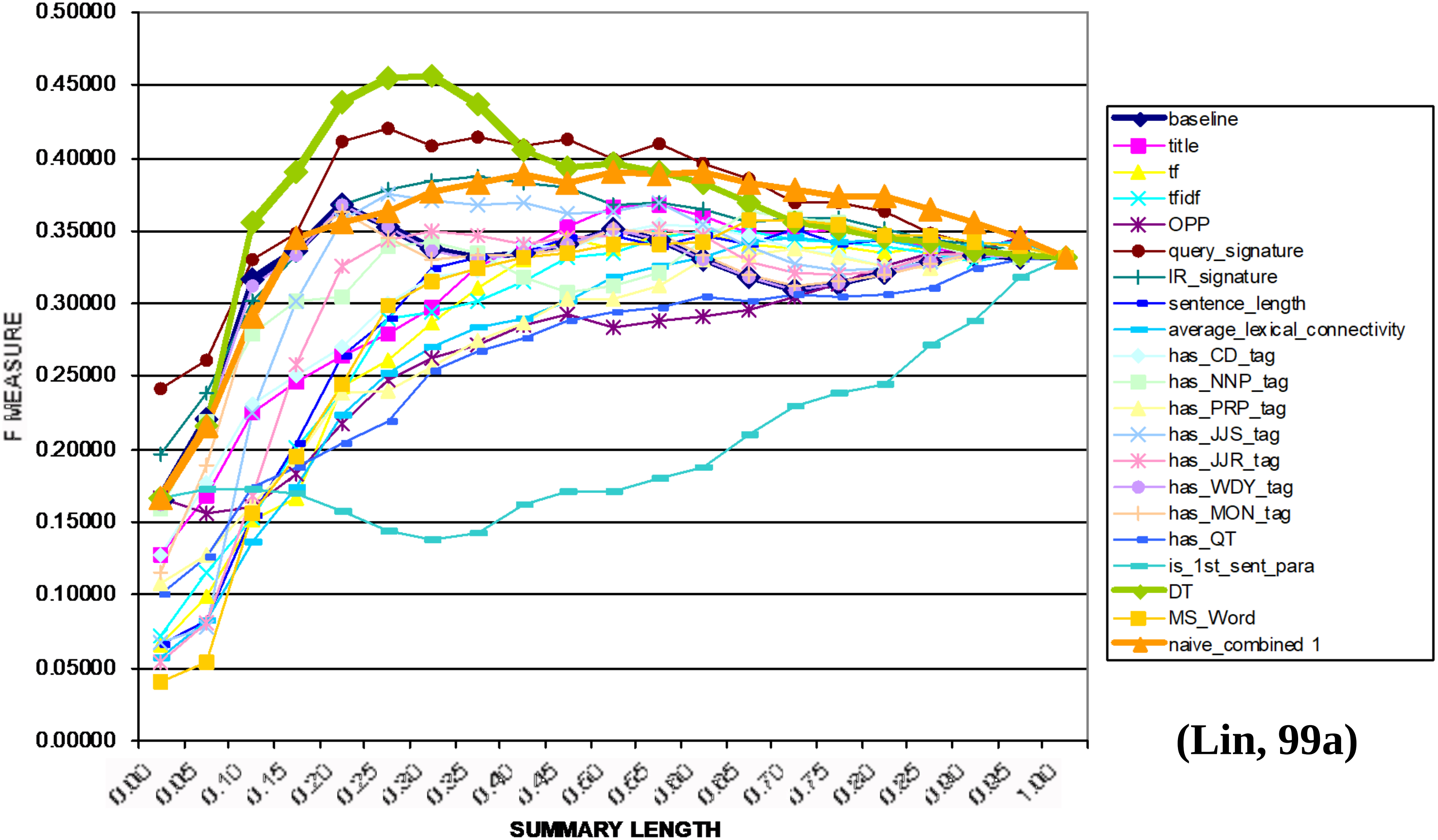
- Information extraction forms
- Query-driven extraction:
  - query expansion lists
  - co-reference with query names
  - lexical similarity to query



Finally: Combining the Evidence

- **Problem**: which extraction methods to use?
- **Answer**: assume they are independent, and combine their evidence: merge individual sentence scores.
- **Studies**:
  - (Kupiec et al., 95; Aone et al., 97, Teufel and Moens, 97): Bayes' Rule.
  - (Mani and Bloedorn,98): SCDF, C4.5, inductive learning.
  - (Lin, 99): C4.5, neural network.
  - (Marcu, 2000): rhetorical parsing tuning.

Performance of Individual Factors



(Lin, 99a)

And Now, an Example...

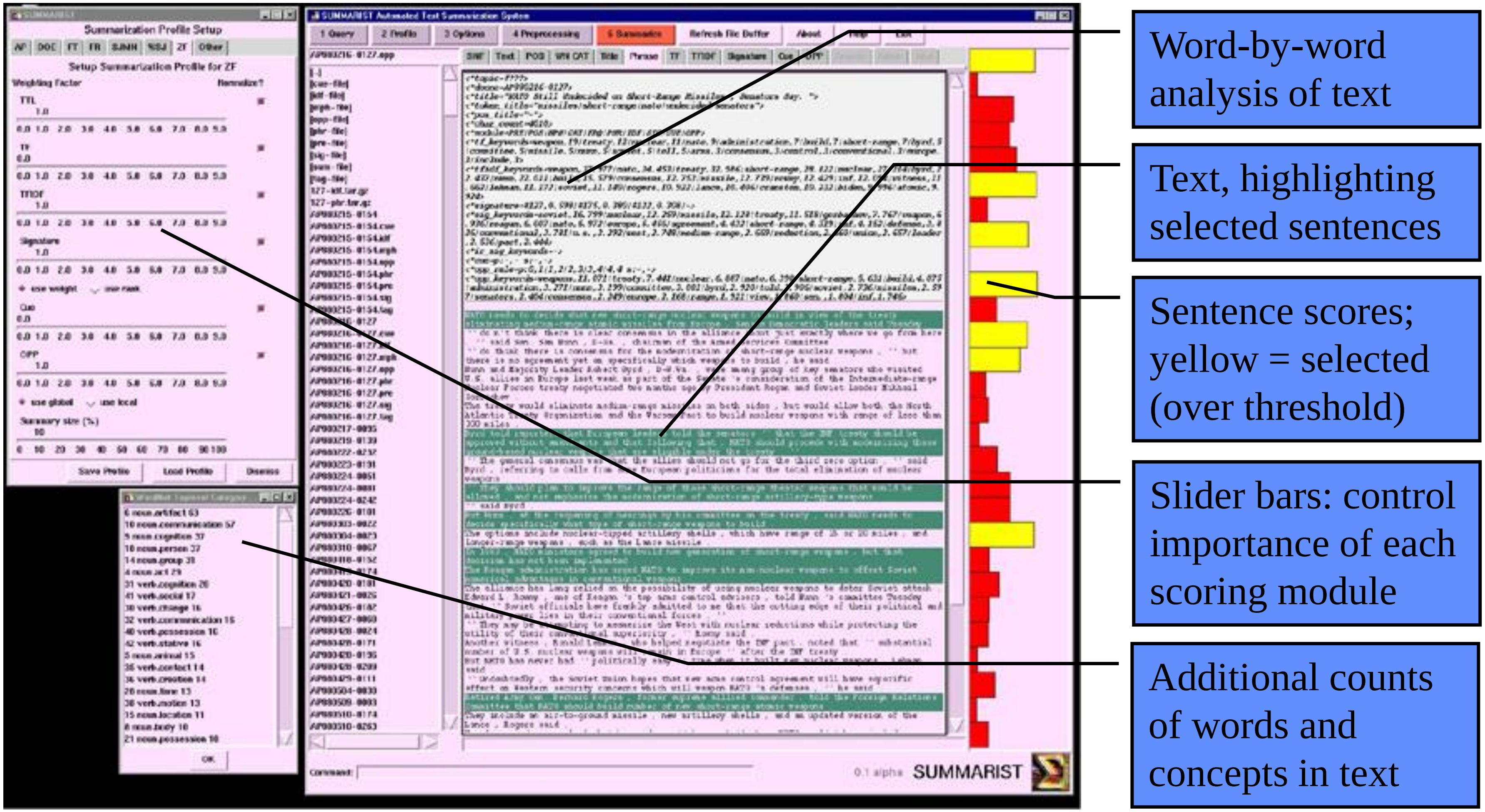
Example System: **SUMMARIST**

Three stages: (Hovy and Lin, 99a; 99b)

SUMMARY = TOPIC ID + INTERPRETATION + GENERATION

1. **Topic Identification Modules**: Positional Importance, Cue Phrases (under construction), Word Counts, Discourse Structure (under construction), ...
2. **Topic Interpretation Modules**: Concept Counting /Wavefront, Concept Signatures (being extended)
3. **Summary Generation Modules** (not yet built): Keywords, Template Gen, Sent. Planner & Realizer

SUMMARIST: Developer's Interface



Internal Format: Preamble

<*docno = AP890417-0167>

<*title = "Former Hostage Accuses Britain of Weakness .">

<*module = PRE|POS|MPH|FRQ|IDF|SIG|CUE|OPP>

<*freq = 544,471,253>

<*tfidf_keywords =
france,13.816|holding,9.210|hostage,8.613|iranian,8.342|television,8.342|writer,7.927|release,7.532|negotiate,7.395|germany, ...>

<*signature = #4,0.577|#2,0.455|#6,0.387>

<*sig_keywords =
hostage,0.725|hold,0.725|western,0.725|moslem,0.725|iranian,0.725|release,0.725|middle,0.725|kill,0.725|west,0.725|march,0.725|east,0.725|syrian, ...>

<*opp_rule = p:0,1|1,2|2,3|3,4|4,4 s:-,->

<*opp_keywords =
kauffmann,4.578|release,3.866|britain,3.811|mccarthy,3.594|hostages,3.406|british,3.150|hostage,2.445|french,2.164|negotiate,2.161| ...>

Internal Format: Word-by-Word

Former <pno=1 sno=1 pos=JJ cwd=1 mph=- frq=1 tfidf=0.000 sig=-,-|-,-|-,- cue=0,- opp=-,->

hostage <pno=1 sno=1 pos=NN cwd=0 mph=- frq=6 tfidf=8.613 sig=1,12.169|33,1.370|2,5.791 cue=0,- opp=2.445,0.898>

John-Paul <pno=1 sno=1 pos=NNP cwd=0 mph=- frq=1 tfidf=0.000 sig=-,-|-,-|-,- cue=0,- opp=0.898,0.898>

Kauffmann <pno=1 sno=1 pos=NNP cwd=0 mph=- frq=6 tfidf=0.000 sig=-,-|-,-|-,- cue=0,- opp=4.578,0.898>

on <pno=1 sno=1 pos=IN cwd=1 mph=- frq=4 tfidf=0.000 sig=-,-|-,-|-,- cue=0,- opp=-,->

Monday <pno=1 sno=1 pos=NNP cwd=0 mph=- frq=3 tfidf=0.000 sig=-,-|-,-|-,- cue=0,- opp=2.076,0.898>

urged <pno=1 sno=1 pos=VBD cwd=0 mph=urge frq=1 tfidf=0.000 sig=-,-|-,-|274,0.492 cue=0,- opp=0.898,0.898>

Example Output, with Keywords

<QNUM>138</QNUM>

<DOCNO>AP890417-0167</DOCNO>

<TITLE>Former Hostage Accuses Britain of Weakness </TITLE>

<TEXT>

Former hostage John-Paul Kauffmann on Monday urged Britain to follow the example set by France and West Germany and negotiate the release of its citizens held captive in Lebanon .

Kauffmann said Britain `` has abandoned '' John McCarthy , 32 , a television reporter abducted on his way to Beirut...

Keywords:

western moslem iranian middle kill march east syrian free anderson group palestinian

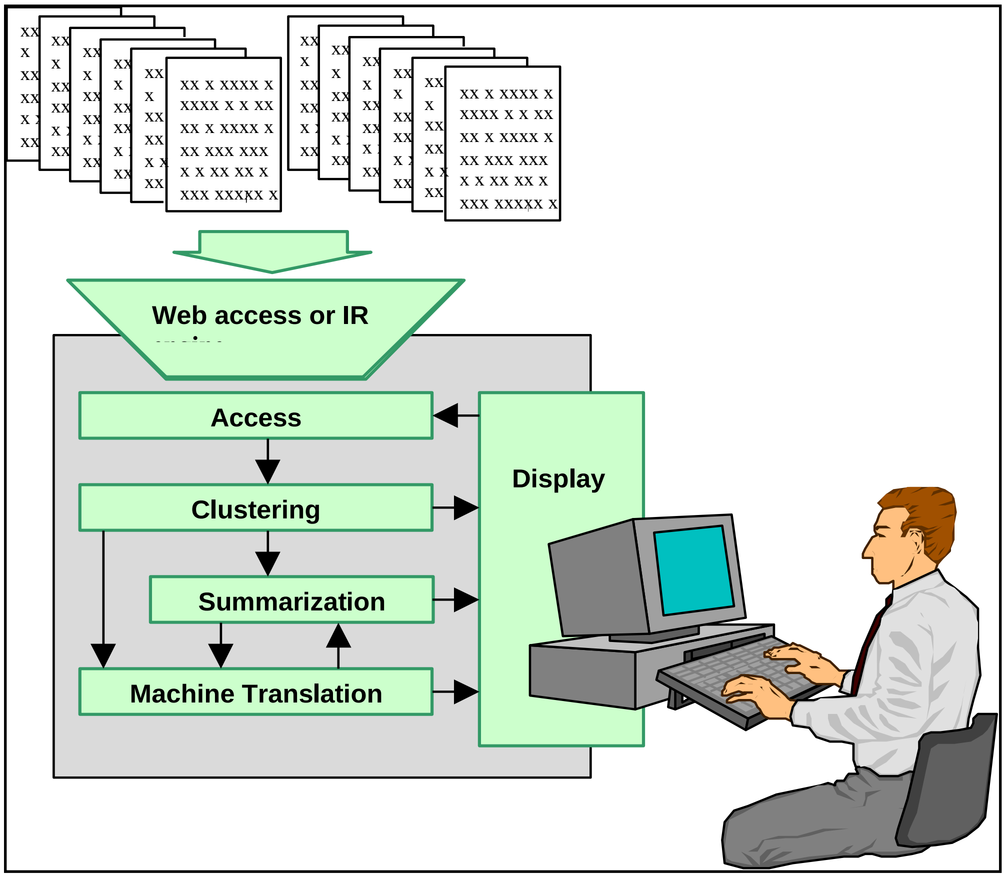
</TEXT>

</DOC>

signature OPP tf.idf

Example System: MuST (Lin and Hovy 98)

- Multilingual Summarization and Translation

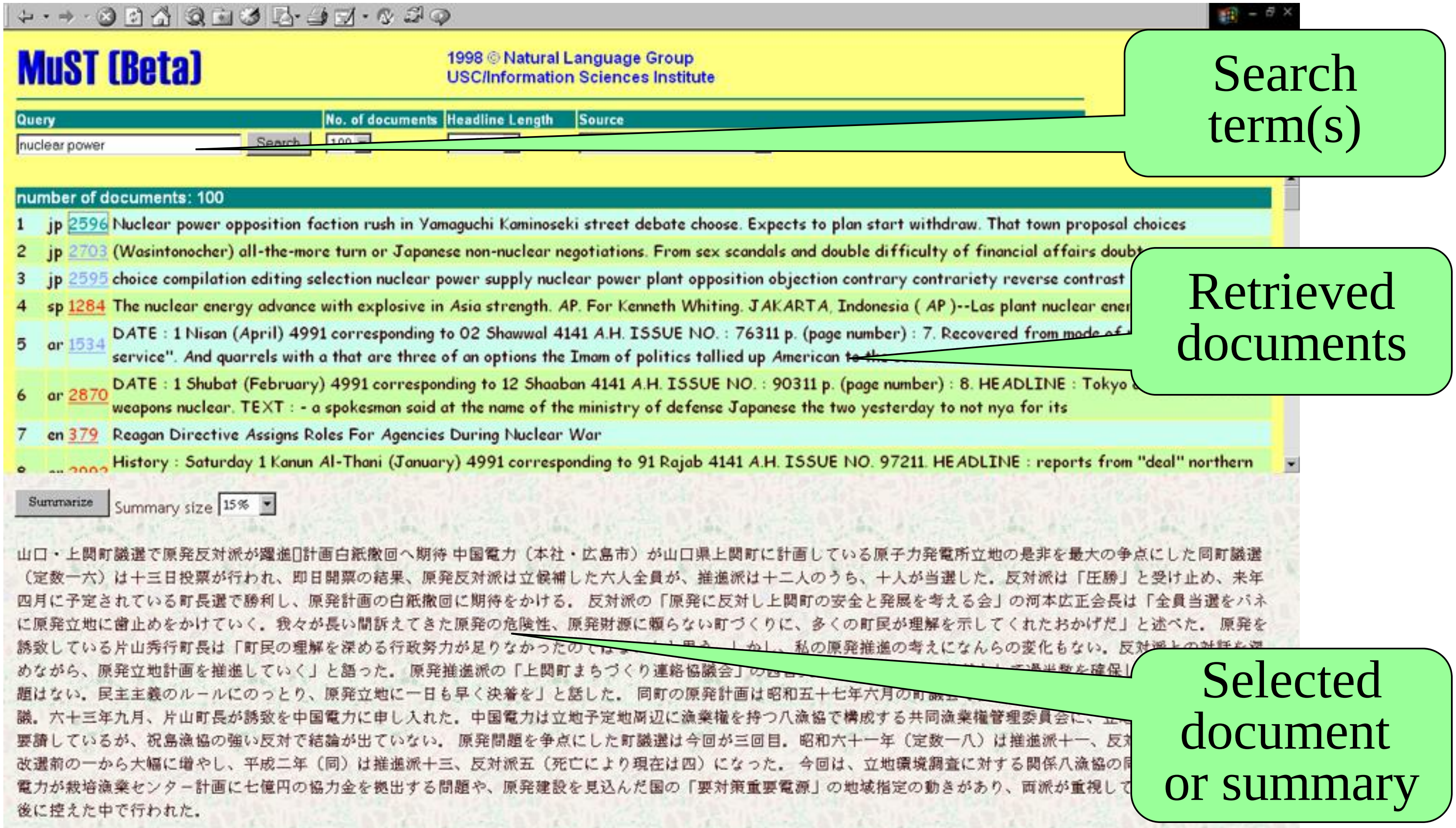


Features:

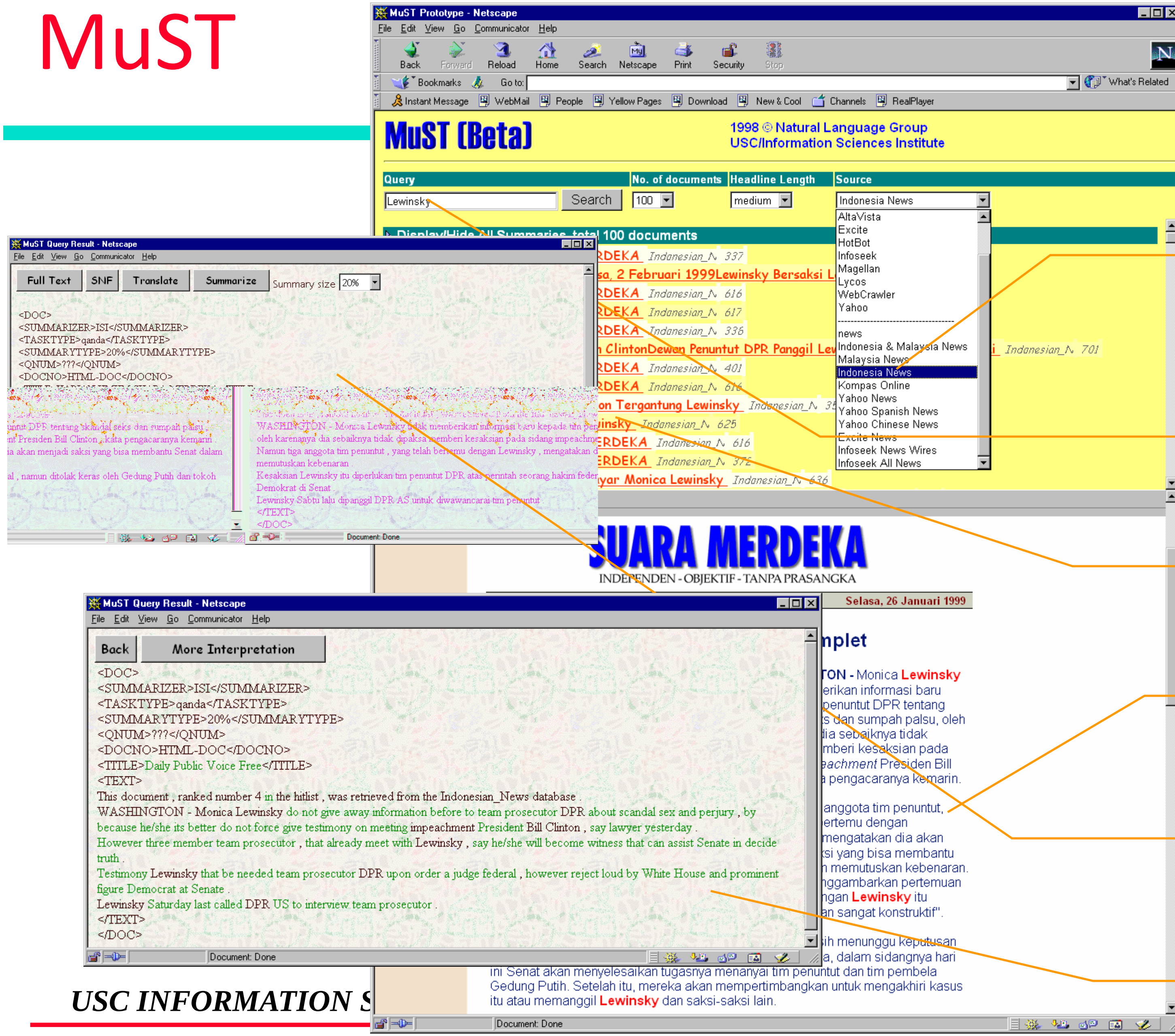
- 8 web search engines
- Local cache for own document collection
- Search and summa-rization of English, Indonesian, Arabic, Spanish, Japanese, (Korean)
- Fast translation of Indonesian; rest slow

MuST Interface

(http://moussor.isi.edu:8080/~cyl/must/must_beta.htm, Lin, 99b)



MuST



Process:

1. Source: web engines or local corpus
2. Search: input terms
3. Retrieved document list
4. Selected document
5. Summary creation
6. Translation

USC INFORMATION S

MuST Translated Web Page


Daily Public <font color="#008800">Voice</ - Netscape

File Edit View Go Communicator Help

Summarize Translate MuST Help

SUARA MERDEKA

INDEPENDEN - OBJEKTIF - TANPA PRASANGKA

International Tuesday, 26 January 1999

Lewinsky Speak Clear and Complete

All together Lawyer: Since arrive at Washington, Monica Lewinsky (middle) continue be in a hurry+ed journalist and fotografer. Lewinsky be accompanied lawyer+s exit from Hotel Mayflower finish eat morning, Monday yesterday.(Photo: *Voice Free* /rtr-gn-52)

WASHINGTON - Monica Lewinsky do not give away information before to team prosecutor DPR about scandal sex and perjury, by because he/she its better do not force give testimony on meeting *impeachment* President Bill Clinton, say lawyer yesterday.

However three member team prosecutor, that already meet with Lewinsky , say he/she will become witness that can assist Senate in decide truth. They draw meeting first with Lewinsky that "productive and very constructive".

Whether Lewinsky called to become witness still wait decision Senate, that will meet again Tuesday this. According to plan, in meeting today Senate will settle task question team prosecutor and team advocate White House. After that, they will consider to end case that or call Lewinsky and witness+s other.

Document: Done

Table of contents

1. Motivation.
2. Genres and types of summaries.
3. Approaches and paradigms.
4. Summarization methods (& exercise).

    Topic Extraction.

    Interpretation.

    Generation.
5. Evaluating summaries.
6. The future.

Summarization exercise

- Write a one-sentence summary for each of the following texts.

Flu stopper
A new compound is set for human testing *(Times)*

Running nose. Raging fever. Aching joints. Splitting headache. Are there any poor souls suffering from the flu this winter who haven't longed for a pill to make it all go away? Relief may be in sight. Researchers at Gilead Sciences a pharmaceutical company in Foster City, California, reported last week in the Journal of the American Chemical Society that they have discovered a compound that can stop the influenza virus from spreading in animals. Tests on humans are set for later this year.

The new compound takes a novel approach to the familiar flu virus. It targets an enzyme, called neuraminidase, that the virus needs in order to scatter copies of itself throughout the body. This enzyme acts like a pair of molecular scissors that slices through the protective mucous linings of the nose and throat. After the virus infects the cells of the respiratory system and begins replicating, neuraminidase cuts the newly formed copies free to invade other cells. By blocking this enzyme, the new compound, dubbed GS 4104, prevents the infection from spreading.

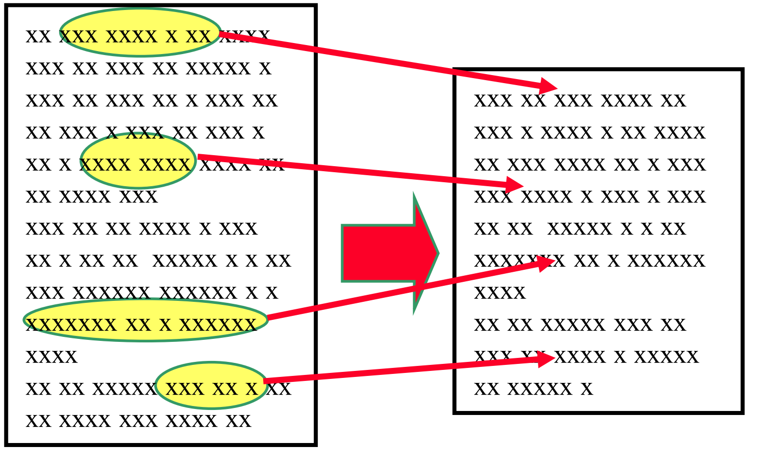
Plant matters

How do you regulate an herb? *(Scientific American)*

If Harlan Page Hubbard were alive, he might be the president of a dietary supplements company. In the late 19th century Hubbard sold Lydia E. Pinkham's Vegetable Compound for kidney and sexual problems. The renowned huckster is remembered each year by national consumer and health organizations who confer a "Hubbard" – a statuette clutching a fresh lemon – for the "most misleading, unfair and irresponsible advertising of the past 12 months."

Appropriately enough, one of this year's winners was a product that Hubbard might have peddled alongside his Lydia Pinkham elixir. Ginkay, an extract of the herb gingko, received its lemon for advertising and labelling claims that someone ingesting the product will have a better memory. Whereas some studies have shown that gingko improves mental functioning in people with dementia, none has proved that it serves as brain tonic for healthy.

Topic Interpretation



- <u>From extract to abstract:</u>
  topic interpretation
  or *concept fusion*.
- <u>Experiment (Marcu, 99):</u>
  - Got 10 newspaper texts, with human abstracts.
  - Asked 14 judges to extract corresponding clauses from texts, to cover the same content.
  - Compared word lengths of extracts to abstracts:
    $$extract\_length = 2.76 \times abstract\_length \;\; !!$$

Some Types of Interpretation

- Concept generalization:

  Sue ate apples, pears, and bananas $\Rightarrow$ Sue ate fruit

- Meronymy replacement:

  Both wheels, the pedals, saddle, chain… $\Rightarrow$ the bike

- Script identification: (Schank and Abelson, 77)

  He sat down, read the menu, ordered, ate, paid, and left $\Rightarrow$ He ate at the restaurant

- Metonymy:

  A spokesperson for the US Government announced that… $\Rightarrow$ Washington announced that...

General Aspects of Interpretation

- <u>Interpretation occurs at the conceptual level...</u>

  …words alone are polysemous (*bat* = *animal* and *sports instrument*) and combine for meaning (*alleged murderer* ≠ *murderer*).

- <u>For interpretation, you need world knowledge...</u>

  …the fusion inferences are not in the text!

- Little work so far: (Lin, 95; Radev and McKeown, 98; Reimer and Hahn, 97; Hovy and Lin, 98).

Form-based operations

- **Claim**: Using IE systems, can aggregate forms by detecting interrelationships.

1. Detect relationships *(contradictions, changes of perspective, additions, refinements, agreements, trends, etc.).*
2. Modify, delete, aggregate forms using rules (Radev and McKeown, 98):

Given two forms,

if (the location of the incident is the same **and**
the time of the first report is before the time of the second report **and**
the report sources are different **and**
at least one slot differs in value)
then combine the forms using a contradiction operator.

Inferences in terminological Logic

- ‘*Condensation*’ *operators* (Reimer and Hahn, 97).

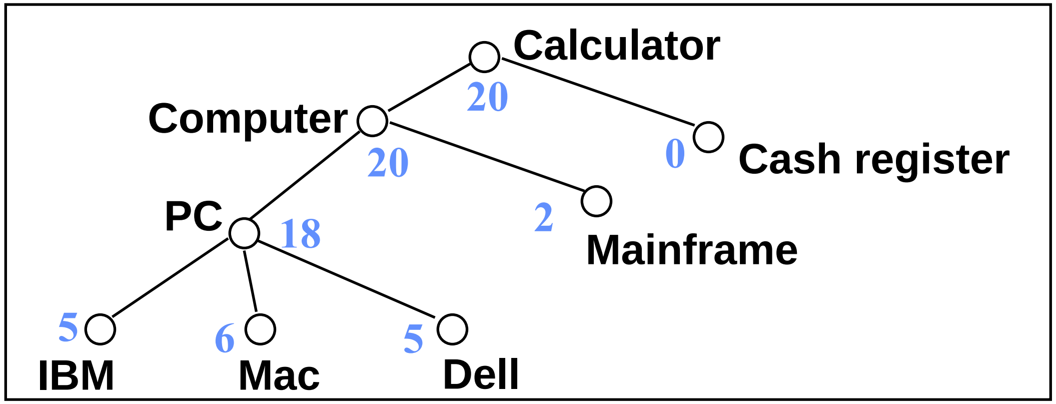
1. Parse text, incrementally build a terminological rep.
2. Apply condensation operators to determine the salient concepts, relationships, and properties for each paragraph (employ frequency counting and other heuristics on concepts and relations, *not* on words).
3. Build a hierarchy of topic descriptions out of salient constructs.

Conclusion: No evaluation.

Concept Generalization: Wavefront

- **Claim**: Can perform concept generalization, using WordNet (Lin, 95).

- Find most appropriate summarizing concept:



1. Count word occurrences in text; score WN concs
2. Propagate scores upward
3. $R = Max\{scores\} / \Sigma\, scores$
4. Move downward until no obvious child: $R<R_t$
5. Output that concept

Wavefront Evaluation

- 200 *BusinessWeek* articles about computers:
  - typical length 750 words (1 page).
  - human abstracts, typical length 150 words (1 par).
  - several parameters; many variations tried.
- $R_t = 0.67$; *StartDepth* = 6; *Length* = 20%:

	Random	Wavefront
Precision	20.30%	33.80%
Recall	15.70%	32.00%

- **Conclusion**: need more elaborate taxonomy.

Topic Signatures (1)

- **Claim**: Can approximate script identification at lexical level, using automatically acquired ‘word families’ (Hovy and Lin, 98).
- **Idea**: Create *topic signatures*: each concept is defined by frequency distribution of its related words (concepts):

  signature = {head (c1,f1) (c2,f2) ...}

  restaurant $\Leftarrow$ *waiter + menu + food + eat...*

- (inverse of query expansion in IR.)

Example Signatures

RANK	aerospace	banking	environment	telecommunication
1	contract	bank	epa	at&t
2	air_force	thrift	waste	network
3	aircraft	banking	environmental	fcc
4	navy	loan	water	cbs
5	army	mr.	ozone	
6	space	deposit	state	bell
7	missile	board	incinerator	long-distance
8	equipment	fslic	agency	telephone
9	mcdonnell	fed	clean	telecommunication
10	northrop	institution	landfill	mci
11	nasa	federal	hazardous	mr.
12	pentagon	fdic	acid_rain	doctrine
13	defense	volcker	standard	service
14	receive	henkel	federal	news
15	boeing	banker	lake	turner
16	shuttle	khoo	garbage	station
17	airbus	asset	pollution	nbc
18	douglas	brunei	city	sprint
19	thiokol	citicorp	law	communication
20	plane	billion	site	broadcasting
21	engine	regulator	air	broadcast
22	million	national_bank	protection	programming
23	aerospace	greenspan	violation	television
24	corp.	financial	management	abc
25	unit	vatican	reagan	rate

Topic Signatures (2)

- **Experiment**: created 30 signatures from 30,000 *Wall Street Journal* texts, 30 categories:
  - Used *tf.idf* to determine uniqueness in category.
  - Collected most frequent 300 words per term.
- **Evaluation**: classified 2204 new texts:
  - Created *document signature* and matched against all topic signatures; selected best match.
- **Results**: *Precision* = 69.31%; *Recall* = 75.66%
  - 90%+ for top 1/3 of categories; rest lower, because less clearly delineated (overlapping signatures).

Evaluating Signature Quality

- Test: perform text categorization task:
  1. match new text's 'signature' against topic signatures
  2. measure how correctly texts are classified by signature
- Document Signature (DS_i):

  $[\ (t_{i1},w_{i1}),\ (t_{i2},w_{i2}),\ \ldots\ ,(t_{in},w_{in})\ ]$
- Similarity measure:
  – cosine similarity, $\cos\theta = TS_k \,.DSi\ \ /\ |\ TS_k\,||DSi|$

DATA	RECALL	PRECISION
7WD	0.847	0.752
7TR	0.844	0.739
7PH	0.843	0.748

DATA	RECALL	PRECISION
8WD	0.803	0.719
8TR	0.802	0.710
8PH	0.797	0.716

Table of contents

NL Generation for Summaries

- **Level 1: no separate generation**
  - Produce extracts, verbatim from input text.
- **Level 2: simple sentences**
  - Assemble portions of extracted clauses together.
- **Level 3: full NLG**
  1. *Sentence Planner*: plan sentence content, sentence length, theme, order of constituents, words chosen... (Hovy and Wanner, 96)
  2. *Surface Realizer*: linearize input grammatically (Elhadad, 92; Knight and Hatzivassiloglou, 95).

Full Generation Example

- **Challenge**: Pack content densely!

- **Example** (Radev and McKeown, 98):
  – Traverse templates and assign values to ‘realization switches’ that control local choices such as tense and voice.
  – Map modified templates into a representation of Functional Descriptions (input representation to Columbia’s NL generation system FUF).
  – FUF maps Functional Descriptions into English.

Generation Example (Radev and mcKeown, 98)

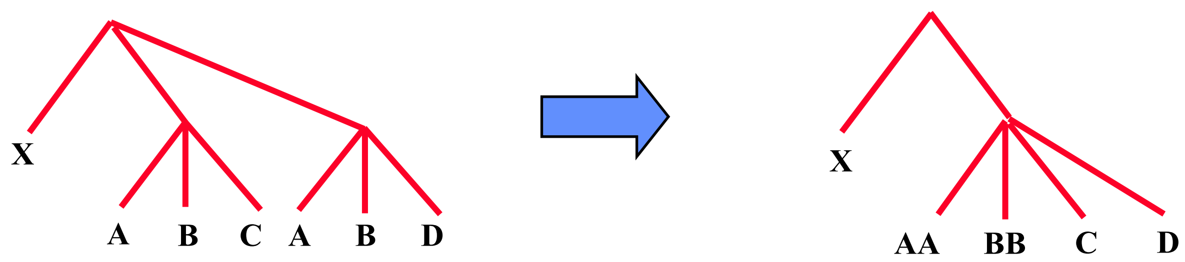
NICOSIA, Cyprus (AP) – Two bombs exploded near government ministries in Baghdad, but there was no immediate word of any casualties, Iraqi dissidents reported Friday. There was no independent confirmation of the claims by the Iraqi National Congress. Iraq's state-controlled media have not mentioned any bombings.

Multiple sources and disagreement

Explicit mentioning of "no information".

Fusion at Syntactic Level

- General Procedure:
  1. Identify sentences with overlapping/related content,
  2. Parse these sentences into syntax trees,
  3. Apply fusion operators to compress the syntax trees,
  4. Generate sentence(s) from the fused tree(s).



Syntax Fusion (1) (Barzilay, McKeown, and Elhadad, 99)

- Parse tree: simple syntactic dependency notation DSYNT, using Collins parser.
- Tree paraphrase rules derived through corpus analysis cover 85% of cases:
  - sentence part reordering,
  - demotion to relative clause,
  - coercion to different syntactic class,
  - change of grammatical feature: tense, number, passive, etc.,
  - change of part of speech,
  - lexical paraphrase using synonym, etc.
- Compact trees mapped into English using FUF.
- Evaluate the fluency of the output.

Syntax Fusion (2) (Mani, Gates, and Bloedorn, 99)

- Elimination of syntactic constituents.
- Aggregation of constituents of two sentences on the basis of referential identity.
- Smoothing:
  - Reduction of coordinated constituents.
  - Reduction of relative clauses.
- Reference adjustment.

Evaluation:

- Informativeness.
- Readability.

Table of contents

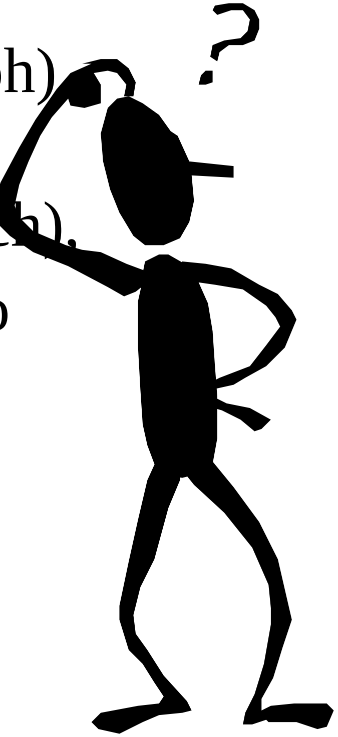
How can You Evaluate a Summary?

- **When you already have a summary…**

  ...then you can compare a new one to it:

  1. choose a granularity (clause; sentence; paragraph),
  2. create a similarity measure for that granularity (word overlap; multi-word overlap, perfect match),
  3. measure the similarity of each unit in the new to the most similar unit(s) in the gold standard,
  4. measure Recall and Precision.

  e.g., (Kupiec et al., 95).



……………….…. but when you don’t?

Two Eval. Types (Sparck Jones and Galliers, 96)

Intrinsic measures (glass-box): how good is the summary as a summary?

- **Problem**: how do you measure the goodness of a summary?
- **Studies**: compare to ideal (Edmundson, 69; Kupiec et al., 95; Salton et al., 97; Marcu, 97) or supply criteria—*fluency*, *informativeness*, *coverage*, etc. (Brandow et al., 95).

Extrinsic measures (black-box): how well does the summary help a user with a task?

- **Problem**: does summary quality correlate with performance?
- **Studies**: GMAT tests (Morris et al., 92); news analysis (Miike et al. 94); IR (Mani and Bloedorn, 97); text categorization (SUMMAC 98; Sundheim, 98).

Extrinsic test: Text Classification

- **<u>Can you perform some task faster?</u>**
  - example: Text Classification.
  - measures: time and effectiveness.
- **TIPSTER/SUMMAC evaluation:**
  - February, 1998 (SUMMAC, 98).
  - Two tests: 1. *Categorization*
    2. *Ad Hoc* (query-sensitive)
  - 2 summaries per system: fixed-length (10%), best.
  - 16 systems (universities, companies; 3 intern'l).

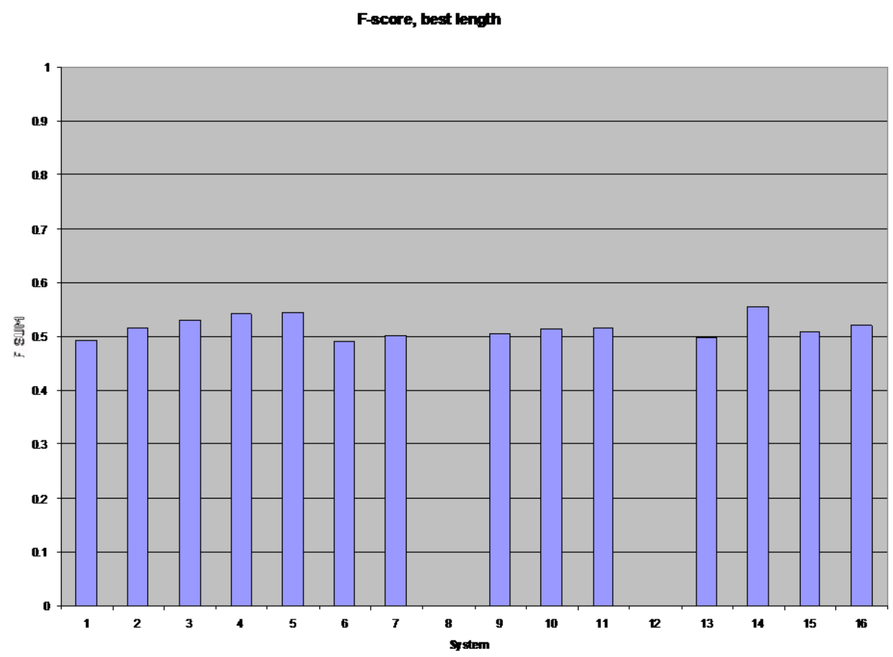
SUMMAC Generic Categorization Test

- **Procedure** (SUMMAC, 98):
  1. 1000 newspaper articles from each of 5 categories.
  2. Systems summarize each text (generic summary).
  3. Humans categorize summaries into 5 categories.
  4. Testers measure *Recall* and *Precision*, combined into *F*: *How correctly are the summaries classified, compared to the full texts?*

  (many other measures as well)

- **Results:**

  No significant difference!

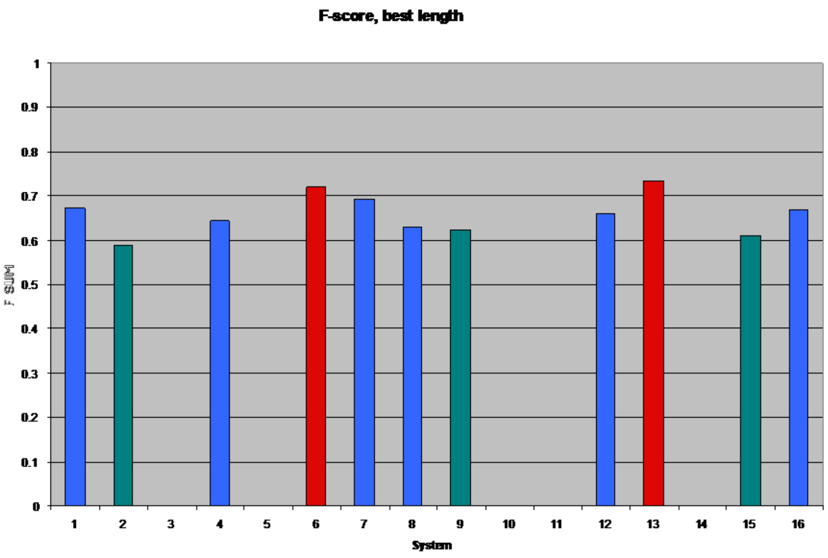


SUMMAC Query-Based 'Ad Hoc' Test

- **Procedure** (SUMMAC, 98):
  1. 1000 newspaper articles from each of 5 categories.
  2. Systems summarize each text (query-based summary).
  3. Humans decide if summary is relevant or not to query.
  4. <u>Testers measure *R* and *P*:</u> *how relevant are the summaries to their queries?* (many other measures as well)

- **Results:**

  3 levels of performance



Intrinsic Test: Q&A Evaluation

- **Can you focus on the important stuff?**

  The Q&A Game—can be tailored to your interests!

- **Measure core info. capture by Q&A game:**
  - Some people (*questioners*) see text, must create questions about most important content.
  - Other people (*answerers*) see:
    1. nothing—but must try to answer questions (baseline),
    2. then: summary, must answer same questions,
    3. then: full text, must answer same questions again.
  - Information retention: % answers correct.

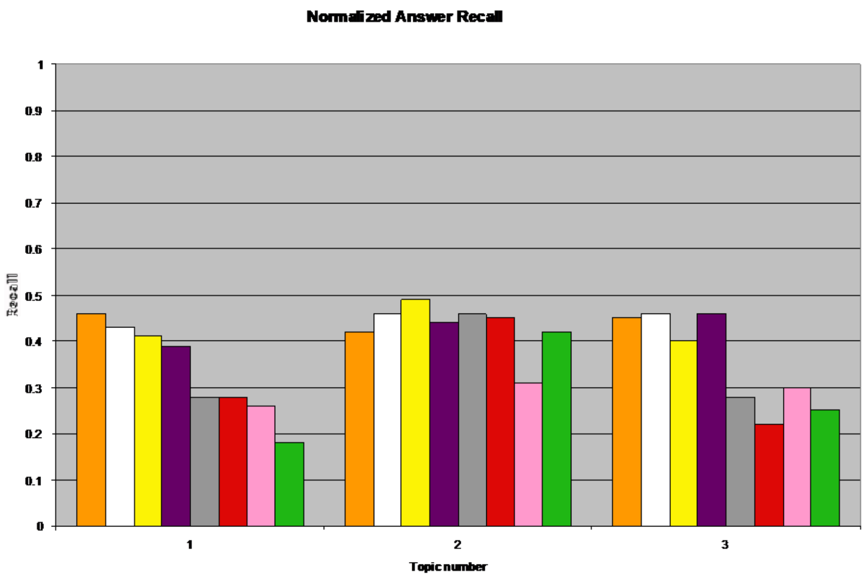
SUMMAC Q&A Evaluation (Sundheim, 98)

- **Procedure** (SUMMAC, 98):
  1. Testers create questions for each category.
  2. Systems create summaries, not knowing questions.
  3. Humans answer questions from originals and from summaries.
  4. <u>Testers measure answer Recall:</u> *how many questions can be answered correctly from the summary?*

     (many other measures as well)

- **Results:**

  Large variation by topic, even within systems...



Comparative Study (Jing et al., 98)

Series of tests on same summaries, to compare different evaluation measures. News genre. 3 systems' summaries scored by 5 judges.

- **Inter-judge agreement is ok**: 96% consistency for news genre for short summaries (10%); 90% consistency for 20% summaries.
- **Summary length is very important**: Precision and Recall vary greatly depending on length, even within single system.

Creating an Ideal Extract (Marcu, 98)

How to compare a human abstract to a system extract?

- Marcu's 'waterfall' method of creating extracts equivalent to abstracts:

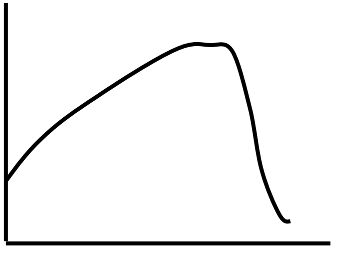
  Given a text and an abstract,

  – determine the text sentence least similar to the abstract (use vector space, word similarity, etc.),

  – discard that sentence, and measure the closeness of the reduced text and abstract,

  – repeat until the closeness starts dropping. Stop.

  – return the remaining text: extract corresponding to the abstract.



- Result: extract-length $\cong$ 2.7 * abstract-length.

Toward a Theory of Evaluation

- **Two Measures:**

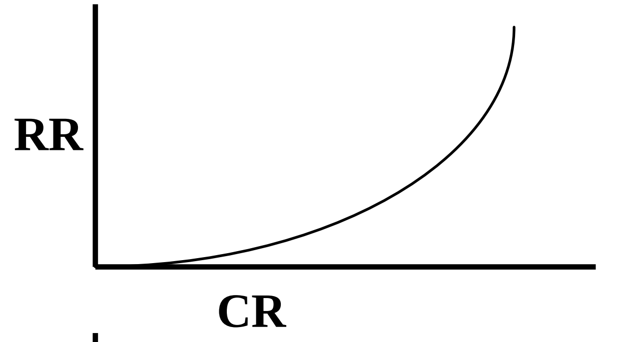
Compression Ratio: *CR = (length S) / (length T)*

Retention Ratio: *RR = (info in S) / (info in T)*

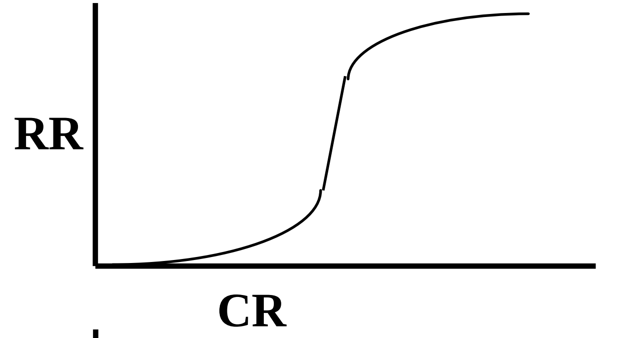
- **Measuring length:**
  - Number of letters? words?
- **Measuring information:**
  - *Shannon Game*: quantify information content.
  - *Question Game*: test reader's understanding.
  - *Classification Game*: compare classifiability.

Compare Length and Information

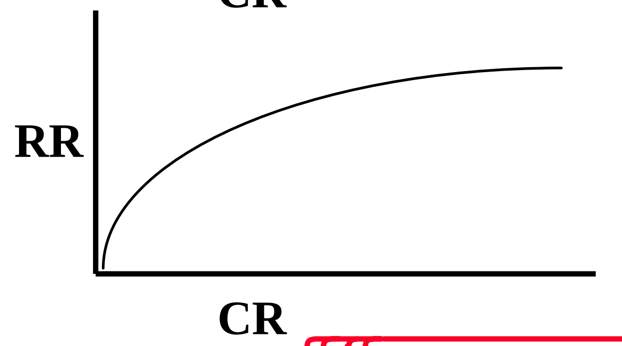
- **Case 1:** just adding info; no special leverage from summary.



- **Case 2:** ‘fuser’ concept(s) at knee add a lot of information.



- **Case 3:** ‘fuser’ concepts become progressively weaker.



Small Eval. Experiment (Hovy and Lin, 99)

- **<u>Can you recreate what's in the original?</u>**
  - the Shannon Game [Shannon 1947–50].
  - but often only *some* of it is really important.
- **Measure info retention (**number of keystrokes**):**
  - 3 groups of subjects, each must recreate text:
    - group 1 sees original text before starting.
    - group 2 sees summary of original text before starting.
    - group 3 sees nothing before starting.
- **Results (# of keystrokes; two different paragraphs):**

Group 1	Group 2	Group 3
approx. 10	approx. 150	approx. 1100

AAAI-98 Symposium Study

- **Burning questions:**
  1. How do <u>different evaluation methods</u> compare for each type of summary?
  2. How do <u>different summary types</u> fare under different methods?
  3. How much does the <u>evaluator</u> affect things?
  4. Is there a preferred evaluation method?
- **Small Experiment**
  - 2 texts, 7 groups.
- **Results:**
  - No difference!
  - As other experiment…
  - ? Extract is best?

		Shannon	Q&A	Classification		
Original		1	1	1	1	1
Abstract	Background	1	3	1	1	1
	Just-the-News		3	1	1	1
	Regular	1	2	1	1	1
Extract	Keywords	2	4	1	1	1
	Random		3	1	1	1
No Text		3	5			
		1-2: 50%	1-2: 30%			
		2-3: 50%	2-3: 20%			
			3-4: 20%			
			4-5:100%			

Table of contents

The Future (1) — **There's much to do!**

- <u>Data preparation:</u>
  - Collect large sets of texts with abstracts, all genres.
  - Build large corpora of <*Text*, *Abstract*, *Extract*> tuples [Marcu, 1999; Jing and McKeown, 1999].
  - Investigate relationships between extracts and abstracts (using <*Extract*, *Abstract*> tuples).
- <u>Types of summary:</u>
  - Determine characteristics of each type.
- <u>Topic Identification:</u>
  - Develop new identification methods (discourse, etc.).
  - Develop heuristics for method combination (train heuristics on <*Text*, *Extract*> tuples).

The Future (2)

- Concept Interpretation (Fusion):
  - Investigate types of fusion (semantic, evaluative…).
  - Create large collections of fusion knowledge/rules (e.g., signature libraries, generalization and partonymic hierarchies, metonymy rules…).
  - Study incorporation of User's knowledge in interpretation.
- Generation:
  - Develop Sentence Planner rules for dense packing of content into sentences (using <*Extract*, *Abstract*> pairs).
- Evaluation:
  - Develop better evaluation metrics, for types of summaries.

Interpretation using Adages

text:

The LA District Attorney has charged Richard Rhee, the owner of a large supermarket chain (California Market) catering to the Asian community, of underreporting more than $4 million in taxes. Rhee, whose preliminary hearing has been set for March 13, faces up to 12 years in prison.

Adages: *Criminal caught and charged*

Roles: Criminal = Richard Rhee, owner of supermarket chain

Crimes = underreporting more than $4 million in taxes

Charger = LA District Attorney

Punishment = up to 12 years in prison

text:

"Shine”, a movie directed by Jane Scott and Scott Hicks, is based on the real-life story of pianist David Helfgott. After being a considerable hit in its native Australia, where it has played for more than 7 months, Hicks had trouble selling it in America. After Miramax co-Chairman Harvey Weinstein agreed to distribute it, the movie grossed over $50 million and won 7 Oscar nominations.

Adages: *Underdog Makes Good* **and** *Persist and you will succeed*

Roles: Underdog = movie "Shine" and makers (Jane Scott, Scott Hicks)

Disbelievers/adversaries = movie studios (Miramax, etc.)

Success = $50 million gross, 7 Oscar nominations, 7 months in Australia

Goodbye!

Appendix

CORPORA IN SUMMARIZATION STUDIES (1)

- Edmundson (68)
  - Training corpus: 200 physical science, life science, information science, and humanities contractor reports.
  - Testing corpus: 200 chemistry contractor reports having lengths between 100 to 3900 words.
- Kupiec et al. (95)
  - 188 scientific/technical documents having an average of 86 sentences each.

Corpora IN summarization studies(2)

- Teufel and Moens (97)
  - 202 computational linguistics papers from the E-PRINT archive.
- Marcu (97)
  - 5 texts from *Scientific American*
- Jing et al. (98)
  - 40 newspaper articles from the TREC collection.
- Marcu (99)
  - 7000 articles from the Ziff-Davis corpus.

CORPORA IN SUMMARIZATION STUDIES(3)

- For each text in each of the first five corpora
  - Human annotators determined the collection of salient sentences/clauses (Edmundson, Jing et al., Marcu97) .
  - One human annotator used author-generated abstracts in order to manually select the sentences that were important in each text (Teufel & Moens).
  - Important sentences were considered to be those that matched closely the sentences of abstracts generated by professional summarizers (Kupiec).

Corpora in summarization studies(4)

- TIPSTER (98)
  - judgments with respect to
    - a query-oriented summary being relevant to the original query;
    - a generic summary being adequate for categorization;
    - a query-oriented summary being adequate to answer a set of questions that pertain to the original query.
- Marcu (99)
  - automatically generated extracts at levels of performance that are close to those of humans.

Making Sense of it All...

To understand summarization, it helps to consider several perspectives simultaneously:

1. **Approaches**: basic starting point, angle of attack, core focus question(s): *psycholinguistics*, *text linguistics*, *computation*...

2. **Paradigms**: theoretical stance; methodological preferences: *rules*, *statistics*, *NLP*, *Info Retrieval*, *AI*...

3. **Methods**: the nuts and bolts: modules, algorithms, processing: *word frequency*, *sentence position*, *concept generalization*...

Query-Driven vs. Text-DRIVEN Focus

- **<u>Top-down: Query-driven focus</u>**
  - *Criteria of interest* encoded as search specs.
  - System uses specs to filter or analyze text portions.
  - <u>Examples</u>: *templates* with slots with semantic characteristics; *termlists* of important terms.
- **<u>Bottom-up: Text-driven focus</u>**
  - *Generic importance metrics* encoded as strategies.
  - System applies strategies over rep of whole text.
  - <u>Examples</u>: degree of *connectedness* in semantic graphs; *frequency* of occurrence of tokens.

References (1)

Aone, C., M.E. Okurowski, J. Gorlinsky, B. Larsen. 1997. A Scalable Summarization System using Robust NLP. *Proceedings of the Workshop on Intelligent Scalable Text Summarization*, 66–73. ACL/EACL Conference, Madrid, Spain.

Baldwin, B. and T. Morton. 1998. Coreference-Based Summarization. In T. Firmin Hand and B. Sundheim (eds). TIPSTER-SUMMAC Summarization Evaluation. *Proceedings of the TIPSTER Text Phase III Workshop*. Washington.

Barzilay, R. and M. Elhadad. 1997. Using Lexical Chains for Text Summarization. In *Proceedings of the Workshop on Intelligent Scalable Text Summarization* at the ACL/EACL Conference, 10–17. Madrid, Spain.

Barzilay, R., K. McKeown, and M. Elhadad. 1999. Information Fusion in the Context of Multi-Document Summarization. In *Proceedings of the ACL Meeting*, University of Maryland.

Baxendale, P.B. 1958. Machine-Made Index for Technical Literature—An Experiment. *IBM Journal* (October) 354–361.

Boguraev B. and C. Kennedy, 1997. Salience-based Content Characterization of Text Documents. In *Proceedings of the Workshop on Intelligent Scalable Text Summarization* at the ACL/EACL Conference, 2–9. Madrid, Spain.

Buckley, C. and C. Cardie. 1997. SMART Summarization System. In T. Firmin Hand and B. Sundheim (eds). TIPSTER-SUMMAC Summarization Evaluation. *Proceedings of the TIPSTER Text Phase III Workshop*. Washington.

DeJong, G. 1978. Fast Skimming of News Stories: The FRUMP System. Ph.D. diss. Yale University.

Donlan, D. 1980. Locating Main Ideas in History Textbooks. *Journal of Reading*, 24, 135–140.

Edmundson, H.P. 1968. New Methods in Automatic Extraction. *Journal of the ACM* 16(2), 264–285.

Elhadad, M. 1992. Using Argumentation to Control Lexical Choice: A Functional Unification-Based Approach. Ph.D. diss, Columbia University.

Endres-Niggemeyer, B. 1998. Summarizing Information. New York: Springer-Verlag.

Hovy, E.H. and L. Wanner. 1996. Managing Sentence Planning Requirements. In *Proceedings of the Workshop on Gaps and Bridges in NL Planning and Generation*, 53–58. ECAI Conference. Budapest, Hungary.

Hovy, E.H. and Lin, C-Y. 1999. Automated Text Summarization in SUMMARIST. In M. Maybury and I. Mani (eds), *Intelligent Scalable Summarization Text Summarization*.

Hovy, E.H. and Lin, C.-Y. 1999. Automated Multilingual Text Summarization and its Evaluation. Submitted.

References (2)

Jing, H., R. Barzilay, K. McKeown, and M. Elhadad. 1998. Summarization Evaluation Methods: Experiments and Analysis. *In Working Notes of the AAAI'98 Spring Symposium on Intelligent Text Summarization*, 60–68. Stanford, CA.

Jing, H. and K. McKeown 1999. The Decomposition of Human-Written Summary Sentences. *Proceedings of the SIGIR Meeting*, Berkeley, CA.

Kintsch, W. and T.A. van Dijk. 1978. Toward a Model of Text Comprehension and Production. *Psychological Review*, 85, 363–394.

Knight, K. and V. Hatzivassiloglou. 1995. Two-Level Many-Paths Generation. In *Proceedings of the Thirty-third Conference of the Association of Computational Linguistics (ACL-95)*, 252–260. Boston, MA.

Kupiec, J., J. Pedersen, and F. Chen. 1995. A Trainable Document Summarizer. In *Proceedings of the Eighteenth Annual International ACM Conference on Research and Development in Information Retrieval* (SIGIR), 68–73. Seattle, WA.

Lehnert, W.G. 1983. Narrative complexity based on summarization algorithms. In *Proceedings of the Eighth International Joint Conference of Artificial Intelligence (IJCAI*-83), 713–716. Karlsruhe, Germany.

Lin, C-Y. 1995. Topic Identification by Concept Generalization. In *Proceedings of the Thirty-third Conference of the Association of Computational Linguistics (ACL-95)*, 308–310. Boston, MA.

Lin, C-Y. 1997. Robust Automated Topic Identification. Ph.D. diss., University of Southern California.

Lin, C-Y. and E.H. Hovy. 1997. Identifying Topics by Position. In *Proceedings of the Applied Natural Language Processing Conference (ANLP-97),* 283–290. Washington.

Lin, C.-Y. 1999. Training a Selection Function for Extraction in SUMMARIST. In *Proceedings of the Eighteenth Annual International ACM Conference on Information and Knowledge Management* (CIKM), Kansas City, Kansas.

Lin, C.-Y. 1999. Machine Translation for Information Access across the Language Barrier: the MuST System. In *Proceefings of the Machine Translation Summit VII*, Singapore.

Luhn, H.P. 1959. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 159–165.

Mani, I., E. Bloedorn, and B. Gates. 1998. Using Cohesion and Coherence Models for Text Summarization. *In Working Notes of the AAAI'98 Spring Symposium on Intelligent Text Summarization*, 69–76. Stanford, CA.

Mani I. And E. Bloedorn. 1998. Machine Learning of Generic and User-Focused Summarization. *Proceedings of the National Conference on Artificial Intelligence, (AAAI).* Madison, WI.

References (3)

Mani I, B. Gates, and E. Bloedorn. 1999. Improving Summaries by Revising Them. *Proceedings of the National Conference on Artificial Intelligence, (AAAI).* Orlando, FL.

Mann, W.C. and S.A. Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text* 8(3), 243–281. Also available as USC/Information Sciences Institute Research Report RR-87-190.

Marcu, D. 1997. The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts. Ph.D. diss. University of Toronto.

Marcu D. 1999. Discourse Structures Are Good Indicators of Importance in Text. In M. Maybury and I. Mani (eds), *Intelligent Scalable Summarization Text Summarization*. Forthcoming.

Marcu, D. 1999. The Automatic Construction of Large-Scale Corpora for Summarization Research. *Proceedings of the SIGIR Meeting,* Berkeley, CA.

Marcu, D. 2000. *Discourse Processing: From Theoretical and Empirical Linguistics to Applications.* The MIT Press. Forthcoming.

Mauldin, M.L. 1991. *Conceptual Information Retrieval—A Case Study in Adaptive Partial Parsing*. Boston, MA: Kluwer Academic Publishers.

Mitra M., A. Singhal, and C. Buckley. 1997. Automatic Text Summarization by Paragraph Extraction. In *Proceedings of the Workshop on Intelligent Scalable Summarization* at the ACL/EACL Conference, 39–46. Madrid, Spain.

Morris J. and G. Hirst. 1991. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics* 17(1), 21–48.

Morris, A.H., G.M. Kasper, and D.A. Adams. 1992. The Effects and Limitations of Automated Text Condensing on Reading Comprehension. *Information Systems Research* 3(1): 17–35.

MUC conference series. 1989–1997. Sundheim, B. (ed.) *Proceedings of the Message Understanding Conferences, I–VI*. Morgan Kaufman.

Ono K., K. Sumita, and S. Miike. Abstract Generation Based on Rhetorical Structure Extraction. *In Proceedings of the International Conference on Computational Linguistics (Coling)*, 344–348. Japan.

Paice, C.D. 1990. Constructing Literature Abstracts by Computer: Techniques and Prospects. *Information Processing and Management* 26(1): 171–186.

Radev D. and K. McKeown. 1998. Generating Natural Language Summaries from Multiple On-Line Sources. *Computational Linguistics* 24(3), 469-500.

References (4)

Rau, L.S. and P.S. Jacobs. 1991. Creating Segmented Databases from Free Text for Text Retrieval. In *Proceedings of the Fourteenth Annual International ACM Conference on Research and Development in Information Retrieval* (SIGIR), 337–346. New York, NY.

Reimer U. and U. Hahn. 1997. A Formal Model of Text Summarization Based on Condensation Operators of a Terminological Logic. In *Proceedings of the Workshop on Intelligent Scalable Summarization* at the ACL/EACL Conference, 97–104. Madrid, Spain.

Salton, G., J. Allen, C. Buckley, and A. Singhal. 1994. Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts. *Science* 264: 1421–1426.

Schank, R.C. and R.P. Abelson. 1977. *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Sparck Jones, K. and J.R. Galliers. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review*. New York: Springer.

Spark Jones, K. 1997. Invited keynote address, *Workshop on Intelligent Scalable Text Summarization.* ACL/EACL Conference. Madrid, Spain.

SUMMAC, 1998. Firmin Hand, T. and B. Sundheim (eds). TIPSTER-SUMMAC Summarization Evaluation. *Proceedings of the TIPSTER Text Phase III Workshop*. Washington.

Sundheim, B. 1998. TIPSTER-SUMMAC Q&A Summarization Evaluation. *Proceedings of the TIPSTER Text Phase III Workshop*. Washington.

Teufel, S. and M. Moens. 1997. Sentence Extraction as a Classification Task. In *Proceedings of the Workshop on Intelligent Scalable Summarization.* ACL/EACL Conference, 58–65. Madrid, Spain.

Online bibliographies:

- http://www.isi.edu/~cyl/summarization
- http://www.cs.columbia.edu/~radev/summarization/
- http://www.cs.columbia.edu/~jing/summarization.html
- http://www.dcs.shef.ac.uk/~gael/alphalist.html