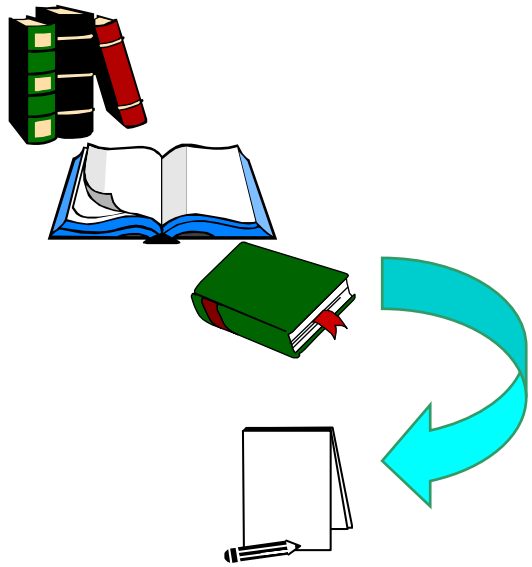
Automated Text Summarization

Stephan Busemann
DFKI GmbH
busemann@dfki.de
http://www.dfki.de/~busemann

Based on the 1998 COLING/ACL Tutorial
by **Ed Hovy** and **Daniel Marcu**, USC-ISI

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

TIME.com | HOME | SEARCH |

TIME Daily
> News Wire
> Editor's Letter
> Comments
> News Features
> Text Only

Magazine
Community
Special Reports

LIFE Picture of the Day

Address

Password

go

Get TIME Daily delivered to your desktop every day with Microsoft Internet Explorer

June 30, 1998

U.S. Plane Fires a Missile On Iraq

An Iraqi radar station targets an Allied plane, and a U.S. F-16 responds quickly -- with deadly force. Is another showdown with Saddam on the way?
Full Story

Responding with Force: A U.S. Air Force F-16 flies over Kuwait. U.S. AIR FORCE/AP

Starr Plays the Tripp Card
The former confidante's grand jury appearance puts the squeeze on Ms. Lewinsky.

Down to Business in Shanghai
President Clinton spends some time in the city he wants the rest of China to turn into.

·Poll: Does the U.S. have the right to impose its idea of human rights on China?

·Postcards From the Middle Kingdom: TIME's Jay Branegan says President Clinton is in full campaign mode in China. But the big question is, why isn't he pressing the flesh?

Boris Duels With the Duma
If Russian president Yeltsin wants to make other Russian pols look bad, he should stop making a fool of himself first.

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

Questions

- What kinds **of summaries do people want?**
  - What are *summarizing, abstracting, gisting*,...?
- How sophisticated **must summarization 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?

Overview

1. Motivation
2. Genres and types of summaries
3. Approaches and paradigms
4. Summarization methods
5. Evaluating summaries

‘Genres’ of Summary?

- **Indicative *vs.* informative**
  ...used for quick categorization vs. content processing.
- **Extract *vs.* abstract**
  ...lists fragments of text vs. re-phrases content coherently.
- **Generic *vs.* query-oriented**
  ...provides author’s view vs. reflects user’s interest.
- **Background *vs.* just-the-news**
  ...assumes reader’s prior knowledge is poor vs. up-to-date.
- **Monolingual vs. cross-lingual**
  ...just summarizes vs. also translates into another language.
- **Single-document *vs.* multi-document source**
  ...based on one text vs. fuses together many texts.

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: amazon customer.

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 **(cf. Sparck Jones 97)**
  - *subject type*: domain
  - *genre*: newspaper articles, editorials, letters, reports...
  - *form*: regular text structure; free-form
  - *source size*: single doc; multiple docs (few; many)
- Purpose
  - *situation*: embedded in larger system (MT, IR) or not?
  - *audience*: focused or general
  - *usage*: IR, sorting, skimming...
- Output
  - *completeness*: include all aspects, or focus on some?
  - *format*: paragraph, table, etc.
  - *style*: informative, indicative, aggregative, critical...
  - *language*: same or other than input

Overview

1. Motivation
2. Genres and types of summaries
3. Approaches and paradigms
4. Summarization methods
5. Evaluating summaries

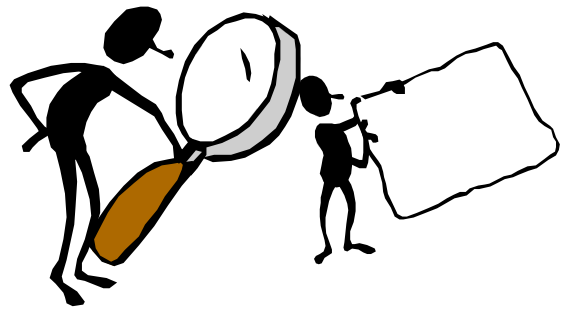
Making Sense of it All...

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

Psycholinguistic Approach: Two 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.

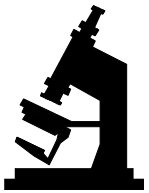


- **552 finely-grained summarization strategies from professional summarizers (Endres-Niggemeyer, 98):**
  - **Self control**: make yourself feel comfortable.
  - **Processing**: produce a unit as soon as you have enough data.
  - **Info organization**: use “Discussion” section to check results.
  - **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 needs: 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 needs: anything that's important**
- **System needs: *generic importance metrics*, used to rate content**

Query-Driven vs. Text-Driven Focus

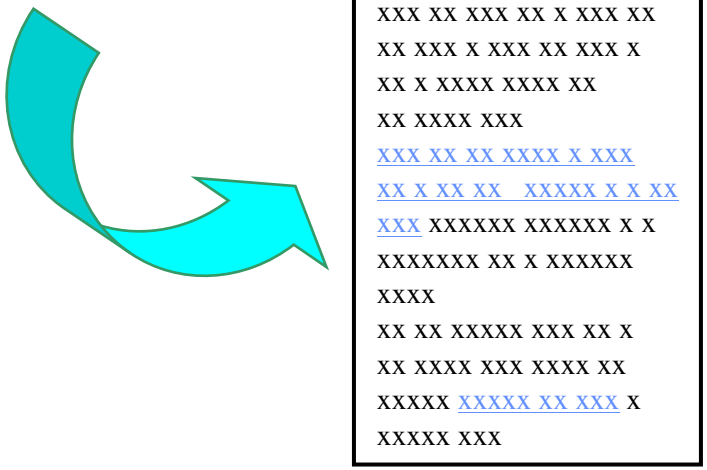
- **Top-down: Query-driven focus**
  - *Criteria of interest* encoded as search specs.
  - System uses specs to filter or analyze text portions.
  - Examples: *templates* with slots with semantic characteristics; *term lists* of important terms.

- **Bottom-up: Text-driven focus**
  - *Generic importance metrics* encoded as strategies.
  - System applies strategies over rep of whole text.
  - Examples: degree of *connectedness* in semantic graphs; *frequency* of occurrence of tokens.

Bottom-Up, using Information Retrieval

- **IR task:** Given a query, find the relevant document(s) from a large set of documents.
- **Summ-IR task:** Given a query, find the relevant passage(s) from a set of passages (i.e., from one or more documents).

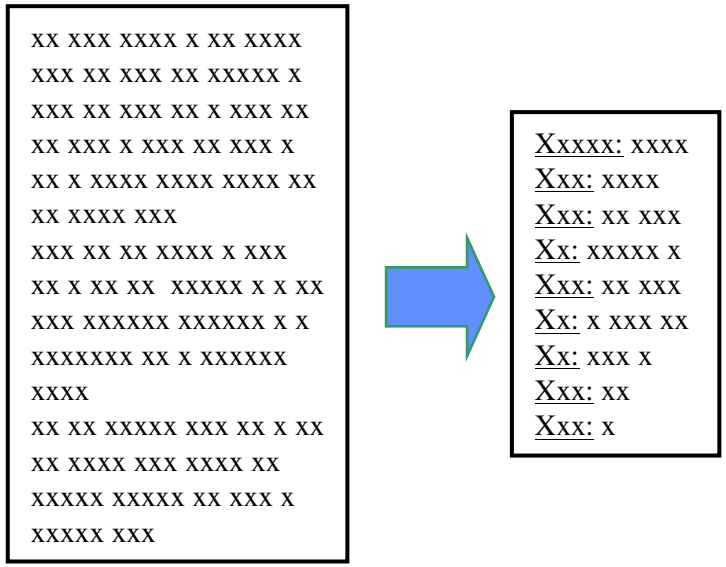
- Questions:
  1. IR techniques work on large volumes of data; can they scale down accurately enough?
  2. IR works on words; do abstracts require abstract representations?



Top-Down, Using Information Extraction

- **IE task:** Given a template and a text, find all the information relevant to each slot of the template and fill it in.
- **Summ-IE task:** Given a query, select the best template, fill it in, and generate the contents.

- Questions:
  1. IE works only for very particular templates; can it scale up?
  2. What about information that doesn't fit into any template—is this a generic limitation of IE?



Paradigms: NLP/IE vs. IR/Statistics

NLP/IE:

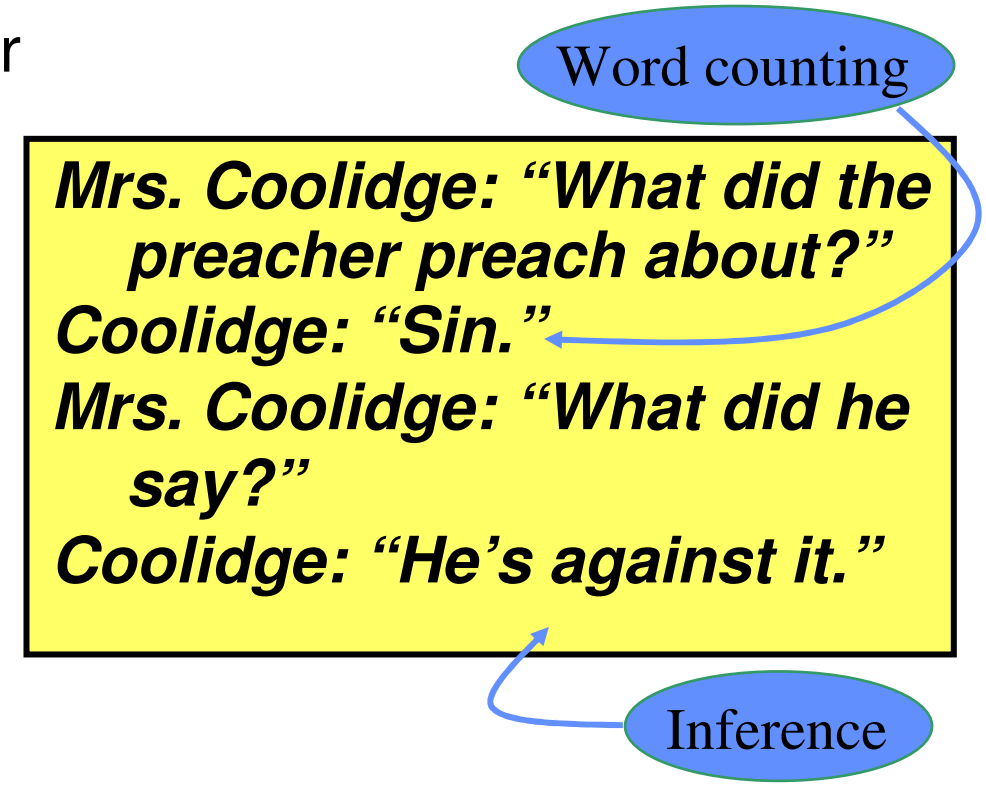
- **Approach**: try to 'understand' text—re-represent content using 'deeper' notation; then manipulate that.
- **Need**: rules for text analysis and manipulation, at all levels.
- **Strengths**: higher quality; supports abstracting.
- **Weaknesses**: speed; still needs to scale up to robust open-domain summarization.

IR/Statistics:

- **Approach**: operate at lexical level—use word frequency, collocation counts, etc.
- **Need**: large amounts of text.
- **Strengths**: robust; good for query-oriented summaries.
- **Weaknesses**: lower quality; inability to manipulate information at abstract levels.

Towards the Final Answer ...

- **Problem:** What if neither IR-like nor IE-like methods work?
  - sometimes counting and templates are insufficient,
  - and then you need to do inference to *understand*.
- **Solution:**
  - semantic analysis of the text (NLP),
  - using adequate knowledge bases that support inference (AI).



The Optimal Solution...

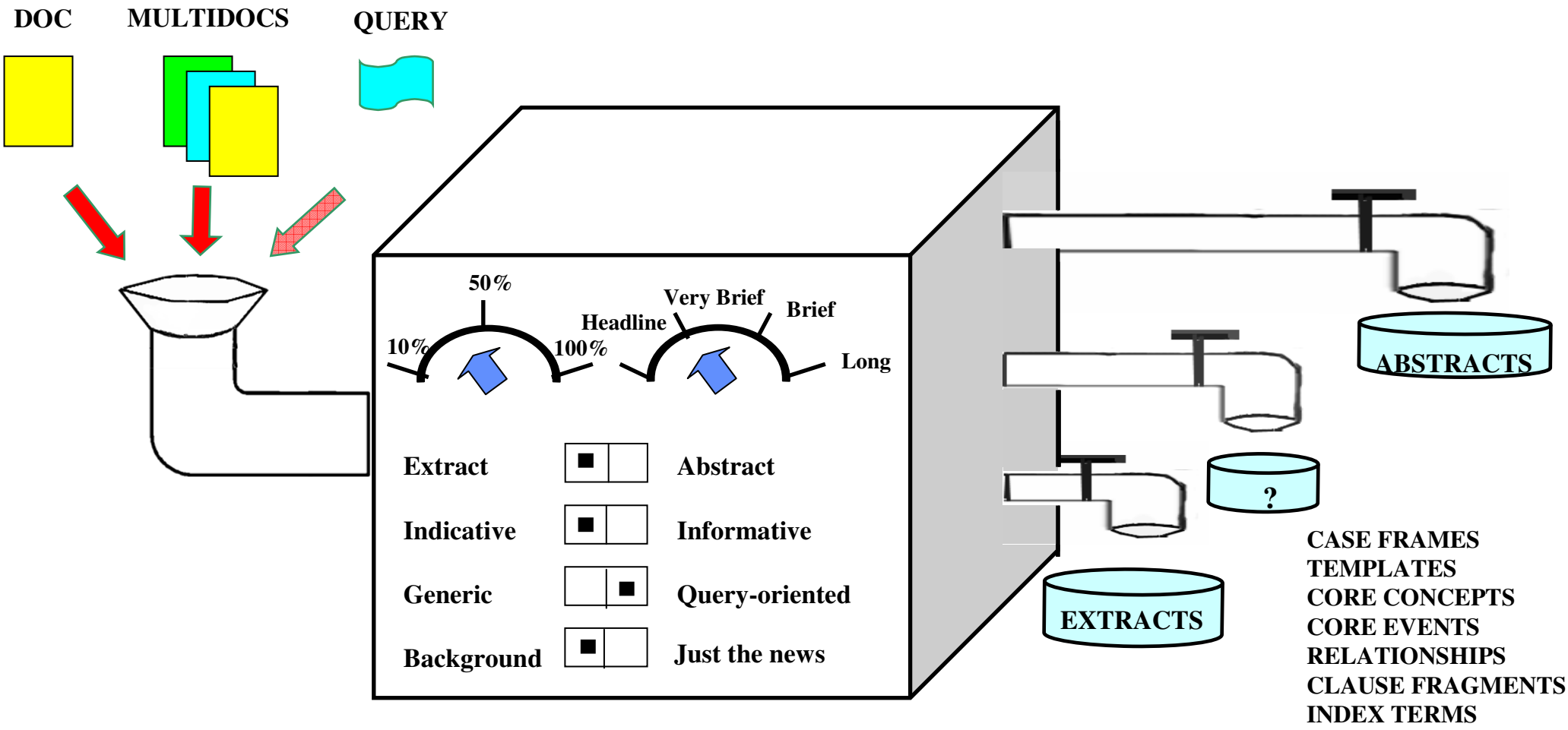
Combine strengths of both paradigms…

...use IE/NLP when you have suitable template(s),

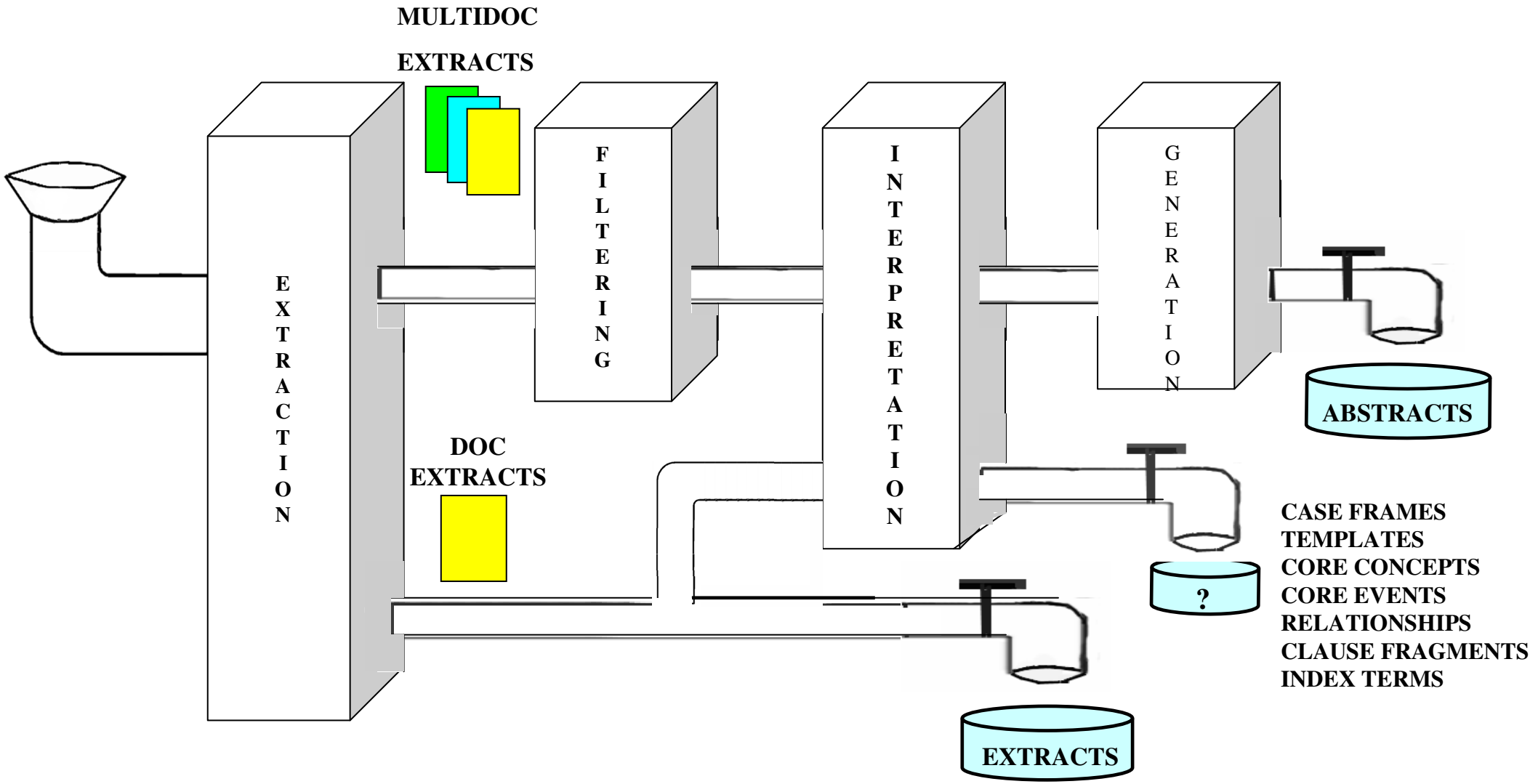
...use IR when you don't…

…but how exactly to do it?

A Summarization Machine



The Modules of the Summarization Machine



Overview

1. Motivation
2. Genres and types of summaries
3. Approaches and paradigms
4. Summarization methods
   - Topic Extraction
   - Interpretation
   - Generation
5. Evaluating summaries

Overview of Extraction Methods

- **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 (1)

- **Claim:** Important sentences are located at positions that are genre-dependent; these positions can be determined automatically through training **(Lin and Hovy, 97).**
  - **Corpus**: 13.000 newspaper articles (ZIFF corpus).
  - **Step 1**: For each article, determine overlap between sentences and the index terms for the article.
  - **Step 2**: Determine a partial ordering over the locations where sentences containing important words occur: Optimal Position Policy (OPP)

Optimum Position Policy (2)

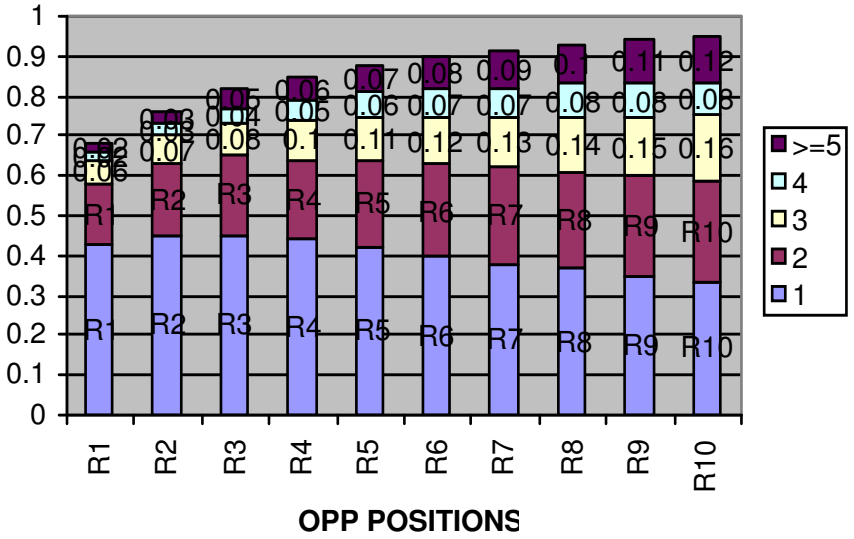
- OPP for ZIFF corpus:

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

  (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.



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)

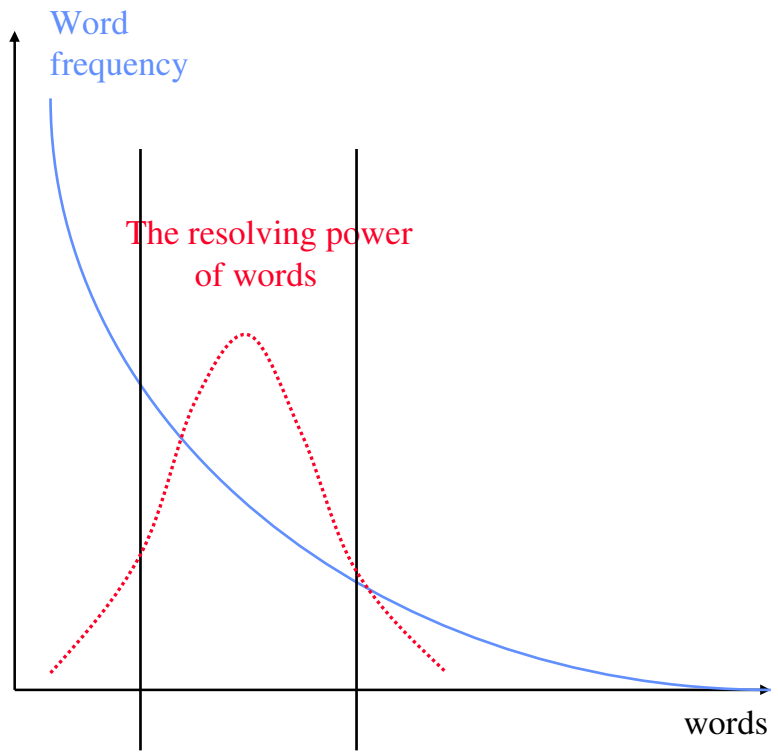
Individual contribution

- **(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.

Word-Frequency-Based Method (1)



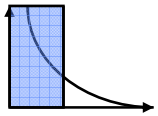
(Luhn, 59)

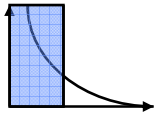
- **Claim:** Important sentences contain words that occur "somewhat" frequently.
- **Method:** Increase sentence score for each frequent word.
- **Evaluation:** Straightforward approach empirically shown to be mostly detrimental in summarization systems.

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)**
  - 17% recall & precision (28% lead baseline)





TF-IDF

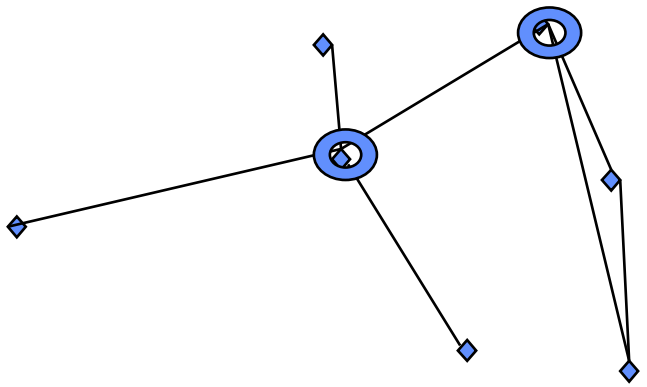
Cumulative contribution

- **(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.

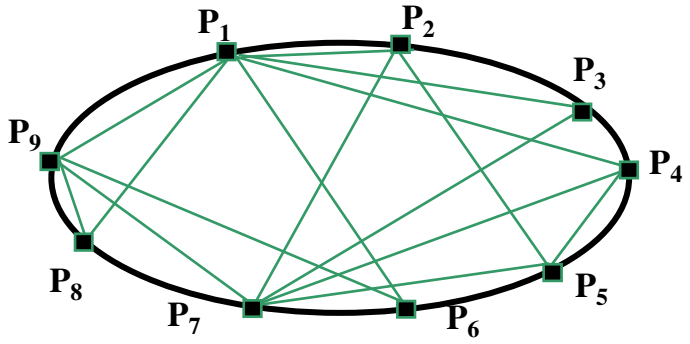
- **Classes of approaches**
  - word co-occurrences;
  - local salience and grammatical relations;
  - co-reference;
  - lexical similarity (WordNet, lexical chains);
  - combinations of the above.



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's Smart-based 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.

    (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.**

Cohesion: 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., 94) for Japanese.
  - (Marcu, 97) for English.

Rhetorical Parsing
(Marcu,97)

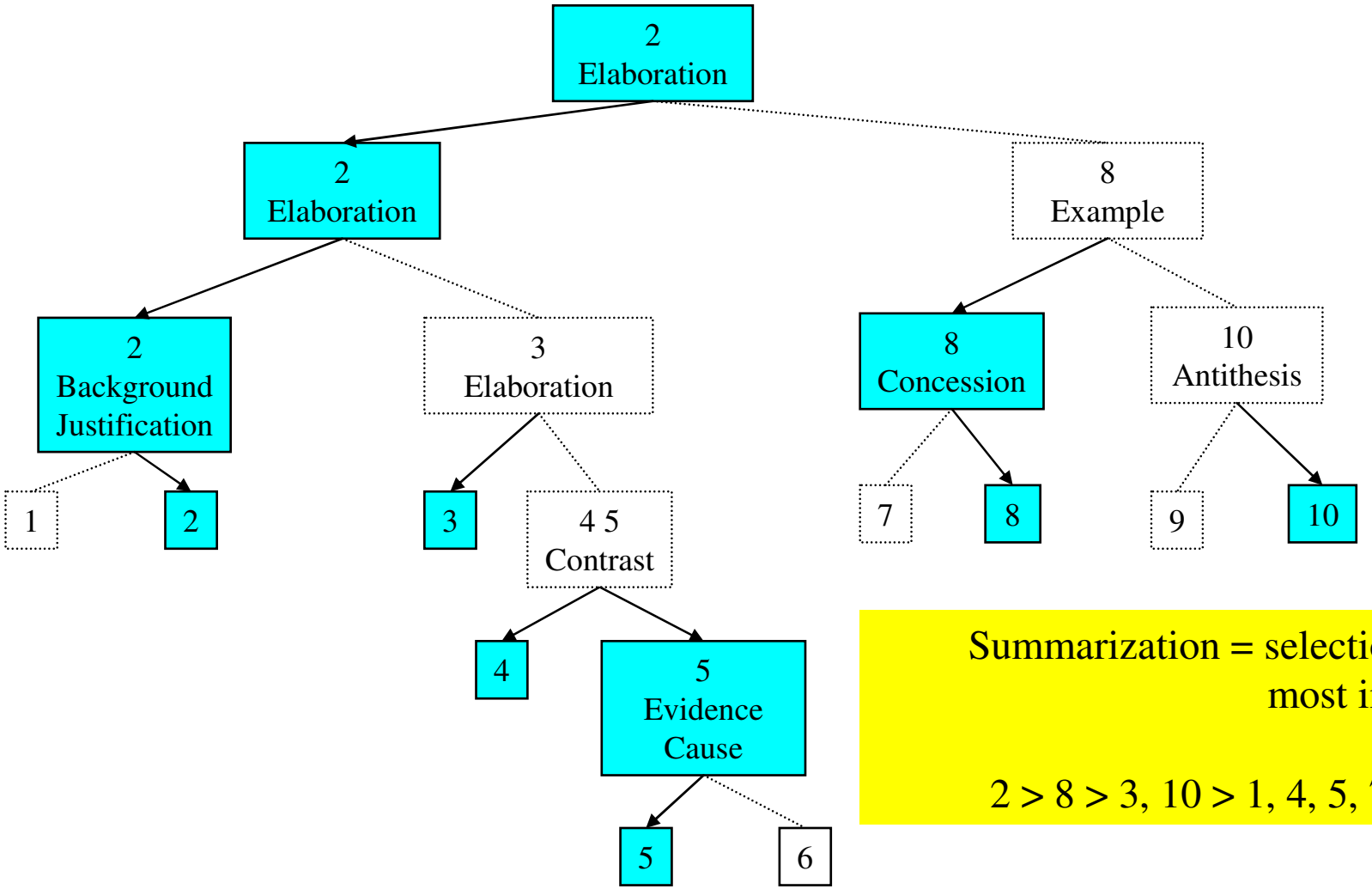
[*With* its distant orbit {– 50 percent farther from the sun than Earth –} and slim atmospheric blanket,[1]] [Mars experiences frigid weather conditions.[2]] [Surface temperatures typically average about –60 degrees Celsius (–76 degrees Fahrenheit) at the equator and can dip to –123 degrees C near the 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(CONTRAST, 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	68.33%	84.16%	75.42%
	Lead	82.91%	63.45%	71.89%
20%	Humans	82.83%	64.93%	72.80%
	Program	59.51%	72.11%	65.21%
	Lead	70.91%	46.96%	56.50%

TREC Corpus

Level	Method	Recall	Precision	F-score
Clause	Humans	72.66%	69.63%	71.27%
	Program	67.57%	73.53%	70.42%
	Lead	39.68%	39.68%	39.68%
Sentence	Humans	78.11%	79.37%	78.73%
	Program	69.23%	64.29%	66.67%
	Lead	54.22%	54.22%	54.22%

Scientific American Corpus

Information Extraction Method (1)

- Idea**: content selection using templates**
  - Predefine a template, 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 template.
  - Generate the content of the template as the summary.
- Previous IE work**:**
  - FRUMP (DeJong, 78): '*sketchy scripts*' of terrorism, natural disasters, political visits...
  - (Mauldin, 91): templates for conceptual IR.
  - (Rau and Jacobs, 91): templates for business.
  - (McKeown and Radev, 95): templates for news.

Information Extraction Method (2)

- **Example template:**

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

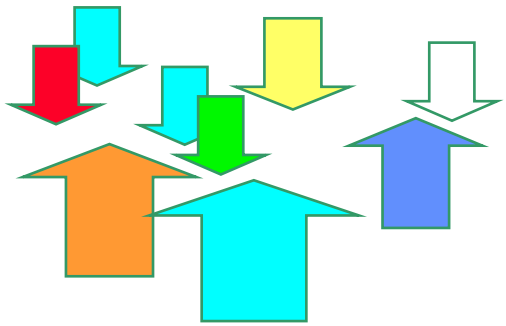
Review of Methods

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 templates**
- **Query-driven extraction:**
  - query expansion lists
  - co-reference with query names
  - lexical similarity to query



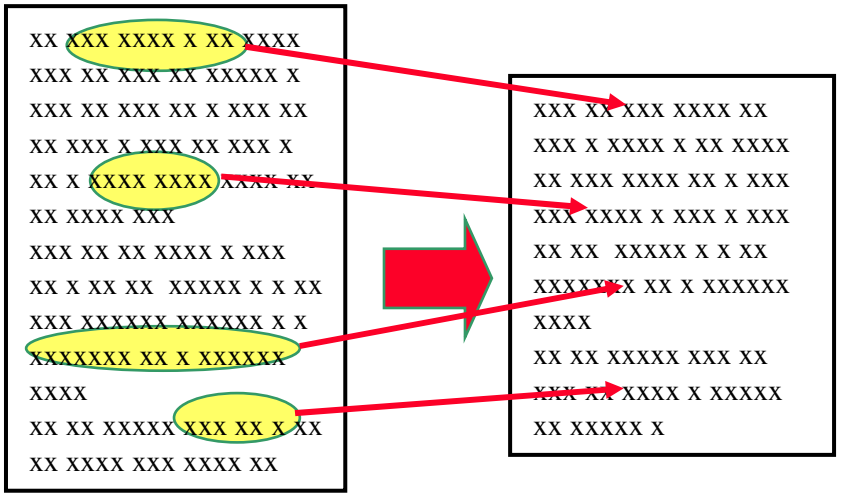
Finally: Combining the Evidence

- **Problem:** which extraction methods to believe?
- **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 and Hovy, 98b): C4.5.
  - (Marcu, 98): rhetorical parsing tuning.

Overview

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

Topic Interpretation



- **From extract to abstract:**
  interpretation

- **Experiment (Marcu, 98):**
  - 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 ⇒ Sue ate fruit

- **Meronymy replacement:**

  Both wheels, the pedals, saddle, chain… ⇒ the bike

- **Script identification:** **(Schank and Abelson, 77)**

  He sat down, read the menu, ordered, ate, paid, and left ⇒
  He ate at the restaurant

- **Metonymy:**

  A spokesperson for the US Government announced that… ⇒
  Washington announced that...

General Aspects of Interpretation

- **Interpretation occurs at the conceptual level...**

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

- **For interpretation, you need world knowledge...**

  …the fusion inferences are not in the text!

Template-based operations

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

1. Detect relationships *(contradictions, changes of perspective, additions, refinements, agreements, trends, etc.).*

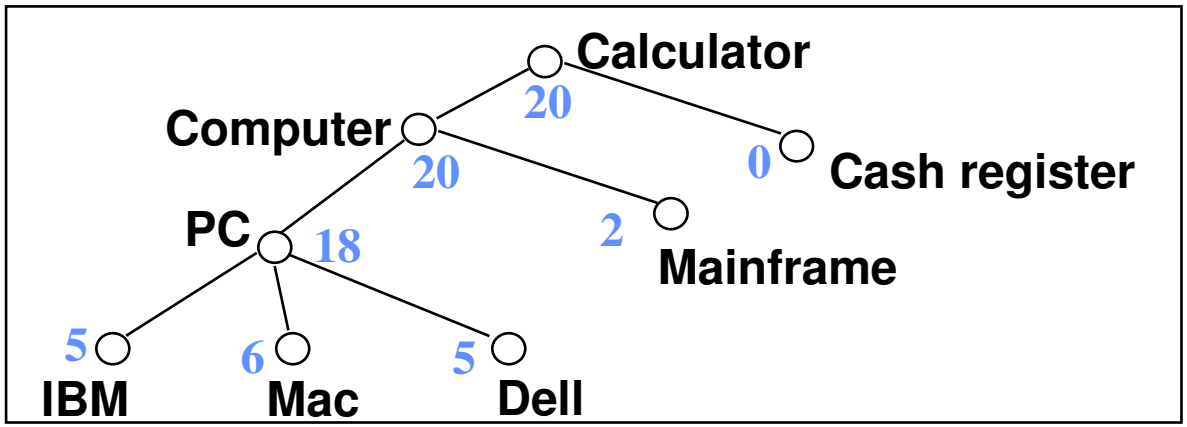
2. Modify, delete, aggregate templates using rules (McKeown and Radev, 95):

Given two templates,
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 templates using a contradiction operator.

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 concepts
2. Propagate scores upward
3. $R = Max\{scores\} / \Sigma scores$
4. Move downward until no obvious child: $R<R_t$ (R_t is a threshold)
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%:**

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

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

Inferences in terminological Logic

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

1. **Parse text, incrementally build a terminological representation**
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**.**

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 ⇐ 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).

Overview

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

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 **(McKeown and Radev, 95):**
  - 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 (McKeown and Radev, 95)

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.

Cross-Lingual Summarization (1)

- **Summary in a language different from that of an input**
- **Needs translation at some stage**
  - Translate as little as necessary, so errors will be minimized
  - Translate as late as possible in the process, so errors won‘t proliferate
- **MUSI: Summarize medical scientific papers in EN and IT into FR and DE**
- **Methods for query-based, indicative summarization in MUSI**
  - Extract sentences using position and cue phrase methods
  - Deeply analyze extracted sentences
  - Re-generate in target language

Cross-Lingual Summarization (2)

(Lenci et al. 2002)

- **Analysis for domain-specific texts (Journal of Anaesthesiology)**
- **Generated text includes optional „meta statements“ about statistics (relevance values)**
- **Performance**
  - better than MT+Summ, worse than Human Summ.
  - MT+Summ scales up better

Language-Neutral Representation for Translating Medical Scientific Text in MUSI

```
PROP{ Value = P_ARG1_cause_ARG2;
      Time_Rep = [PRESENT, PRES_USUAL];
      Cat = V_SEN;
      Arg1 = PROP{ Value = P_antagonism_with_ARG1;
                   Cat = NP; Det = INDEF;
                   Arg1 = ITEM{ Value = C_acetylcholine;
                                Mod1 = [LOC, ITEM{
                                        Value = C_level;
                                        Det = DEF;
                                        Mod1 = [RESTR, ITEM{
                                                Value = C_sight;
                                                Number = PLUR; Det = DEF;
                                                Mod1 = [RESTR, C_muscarinic];
                                                Mod2 = [RESTR, ITEM{
                                                        Value = C_substance;
                                                        Number = PLUR;
                                                        Det = DEMONST1;}]; }]; }]; };
                   Mod1 = [RESTR, C_competitive]; };
      Arg2 = ITEM{ Value = C_effect;
                   Det = DEF; Number = PLUR; }; }
```

The effects are caused by a competitive antagonism with acetylcholine at the level of the muscarinic sights of these substances.

Language-Neutral Representation for Translating Medical Scientific Text in MUSI

```
PROP{ Value = P_ARG1_cause_ARG2;
      Time_Rep = [PRESENT, PRES_USUAL];
      Cat = V_SEN;
      Arg1 = PROP{ Value = P_antagonism_with_ARG1;
                   Cat = NP; Det = INDEF;
                   Arg1 = ITEM{ Value = C_acetylcholine;
                                Mod1 = [LOC, ITEM{
                                        Value = C_level;
                                        Det = DEF;
                                        Mod1 = [RESTR, ITEM{
                                                Value = C_sight;
                                                Number = PLUR; Det = DEF;
                                                Mod1 = [RESTR, C_muscarinic];
                                                Mod2 = [RESTR, ITEM{
                                                        Value = C_substance;
                                                        Number = PLUR;
                                                        Det = DEMONST1;}]; }]; }]; };
                   Mod1 = [RESTR, C_competitive]; };
      Arg2 = ITEM{ Value = C_effect;
                   Det = DEF; Number = PLUR; }; }
```

The effects are caused by a competitive antagonism with acetylcholine at the level of the muscarinic sights of these substances.

Language-Neutral Representation for Translating Medical Scientific Text in MUSI

```
PROP{ Value = P_ARG1_cause_ARG2;
      Time_Rep = [PRESENT, PRES_USUAL];
      Cat = V_SEN;
      Arg1 = PROP{ Value = P_antagonism_with_ARG1;
                   Cat = NP; Det = INDEF;
                   Arg1 = ITEM{ Value = C_acetylcholine;
                                Mod1 = [LOC, ITEM{
                                        Value = C_level;
                                        Det = DEF;
                                        Mod1 = [RESTR, ITEM{
                                                Value = C_sight;
                                                Number = PLUR; Det = DEF;
                                                Mod1 = [RESTR, C_muscarinic];
                                                Mod2 = [RESTR, ITEM{
                                                        Value = C_substance;
                                                        Number = PLUR;
                                                        Det = DEMONST1;}]; }]; }]; };
                   Mod1 = [RESTR, C_competitive]; };
      Arg2 = ITEM{ Value = C_effect;
                   Det = DEF; Number = PLUR; }; }
```

The effects are caused by a competitive antagonism with acetylcholine at the level of the muscarinic sights of these substances.

Language-Neutral Representation for Translating Medical Scientific Text in MUSI

```
PROP{ Value = P_ARG1_cause_ARG2;
      Time_Rep = [PRESENT, PRES_USUAL];
      Cat = V_SEN;
      Arg1 = PROP{ Value = P_antagonism_with_ARG1;
                   Cat = NP; Det = INDEF;
                   Arg1 = ITEM{ Value = C_acetylcholine;
                                Mod1 = [LOC, ITEM{
                                        Value = C_level;
                                        Det = DEF;
                                        Mod1 = [RESTR, ITEM{
                                                Value = C_sight;
                                                Number = PLUR; Det = DEF;
                                                Mod1 = [RESTR, C_muscarinic];
                                                Mod2 = [RESTR, ITEM{
                                                        Value = C_substance;
                                                        Number = PLUR;
                                                        Det = DEMONST1;}]; }]; }]; };
                   Mod1 = [RESTR, C_competitive]; };
      Arg2 = ITEM{ Value = C_effect;
                   Det = DEF; Number = PLUR; }; }
```

The effects are caused by a competitive antagonism with acetylcholine at the level of the muscarinic sights of these substances.

Language-Neutral Representation for Translating Medical Scientific Text in MUSI

```
PROP{ Value = P_ARG1_cause_ARG2;
      Time_Rep = [PRESENT, PRES_USUAL];
      Cat = V_SEN;
      Arg1 = PROP{ Value = P_antagonism_with_ARG1;
                   Cat = NP; Det = INDEF;
                   Arg1 = ITEM{ Value = C_acetylcholine;
                                Mod1 = [LOC, ITEM{
                                        Value = C_level;
                                        Det = DEF;
                                        Mod1 = [RESTR, ITEM{
                                                Value = C_sight;
                                                Number = PLUR; Det = DEF;
                                                Mod1 = [RESTR, C_muscarinic];
                                                Mod2 = [RESTR, ITEM{
                                                        Value = C_substance;
                                                        Number = PLUR;
                                                        Det = DEMONST1;}]; }]; }]; };
                   Mod1 = [RESTR, C_competitive]; };
      Arg2 = ITEM{ Value = C_effect;
                   Det = DEF; Number = PLUR; }; }
```

The effects are caused by a competitive antagonism with acetylcholine at the level of the muscarinic sights of these substances.

Overview

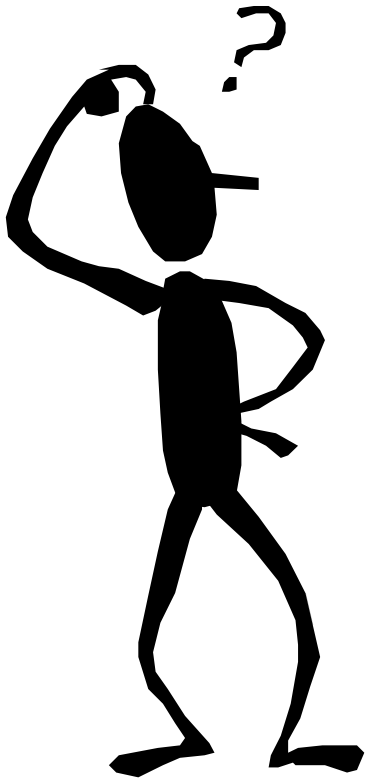
1. Motivation.
2. Genres and types of summaries.
3. Approaches and paradigms.
4. Summarization methods (& exercise).
5. Evaluating summaries.

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?



Toward a Theory of Evaluation

- Two Measures:

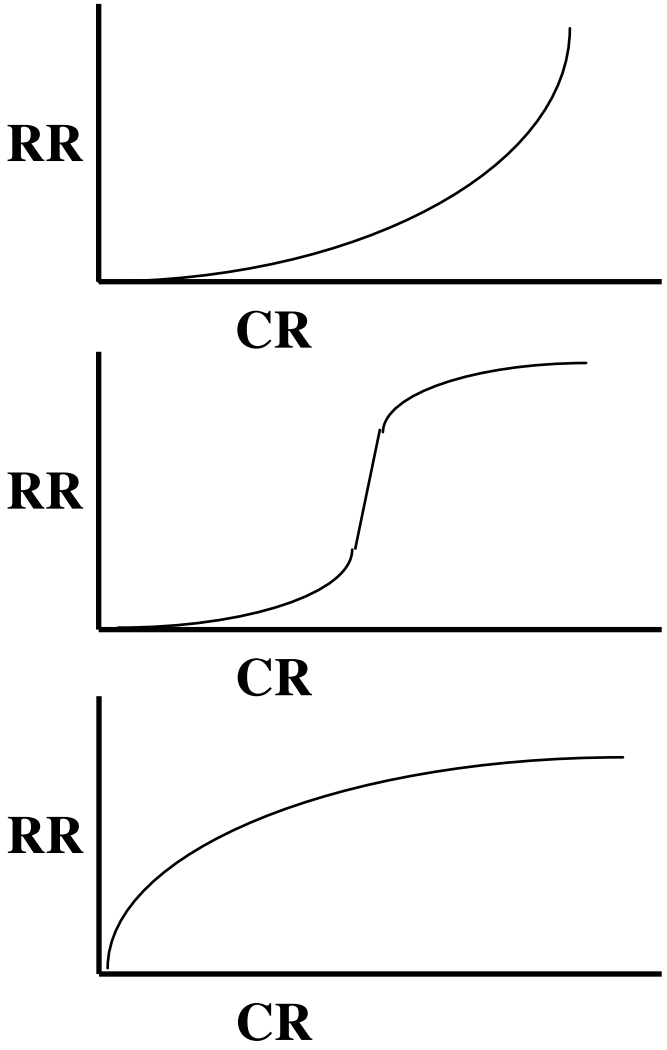
Compression Ratio: $CR = (length\ S) / (length\ T)$

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 Evaluation Experiment
(Hovy, 98)

- Can you recreate what's in the original?
  - 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

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.

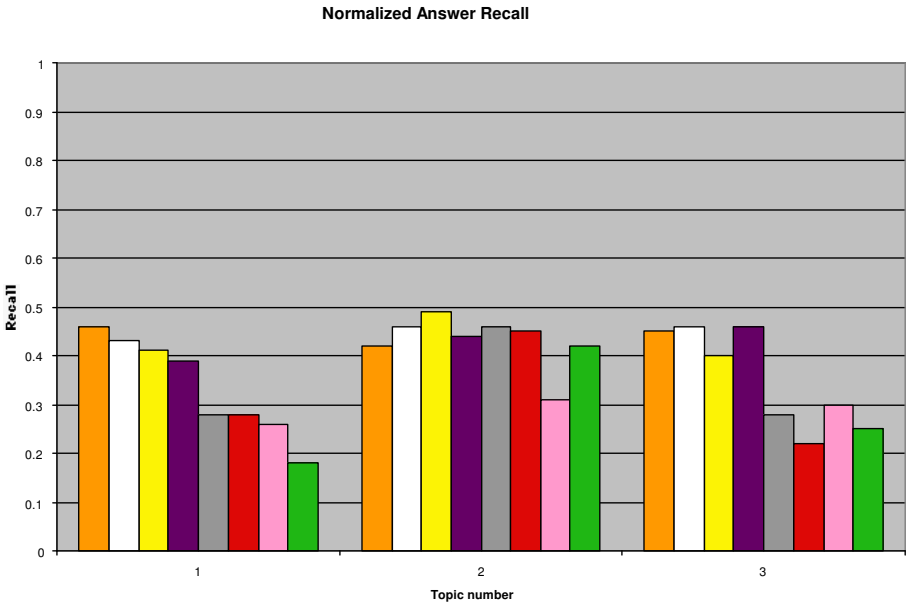
SUMMAC Q&A Evaluation

- 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. Testers measure answer Recall: *how many questions can be answered correctly from the summary?*

     (many other measures as well)

- Results:

  Large variation by topic, even within systems...



Task Evaluation: Text Classification

- Can you perform some task faster?
  - example: the Classification Game.
  - 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).

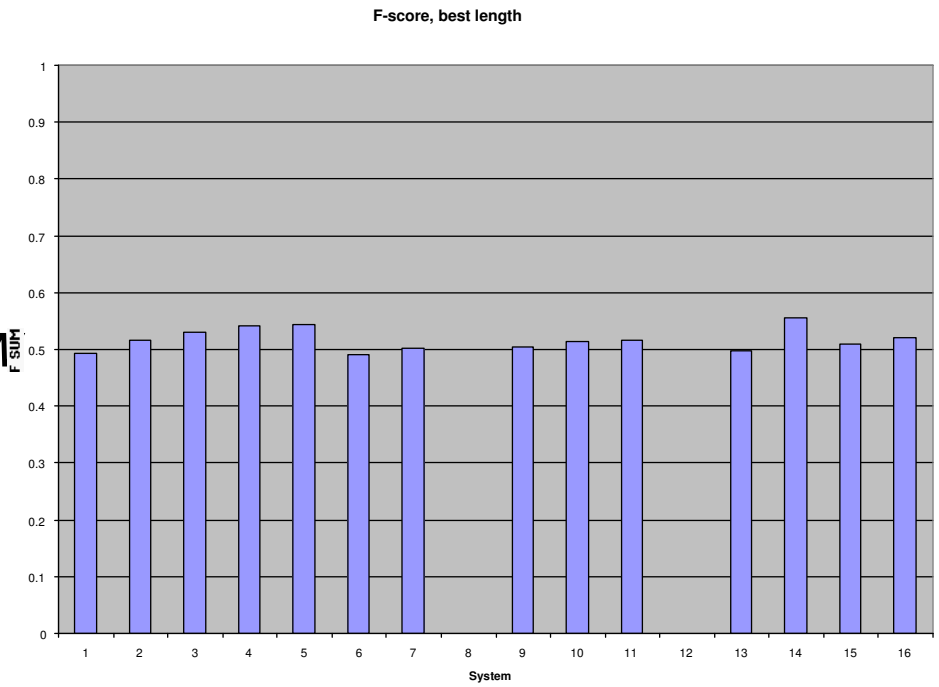
SUMMAC Categorization Test

- 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!



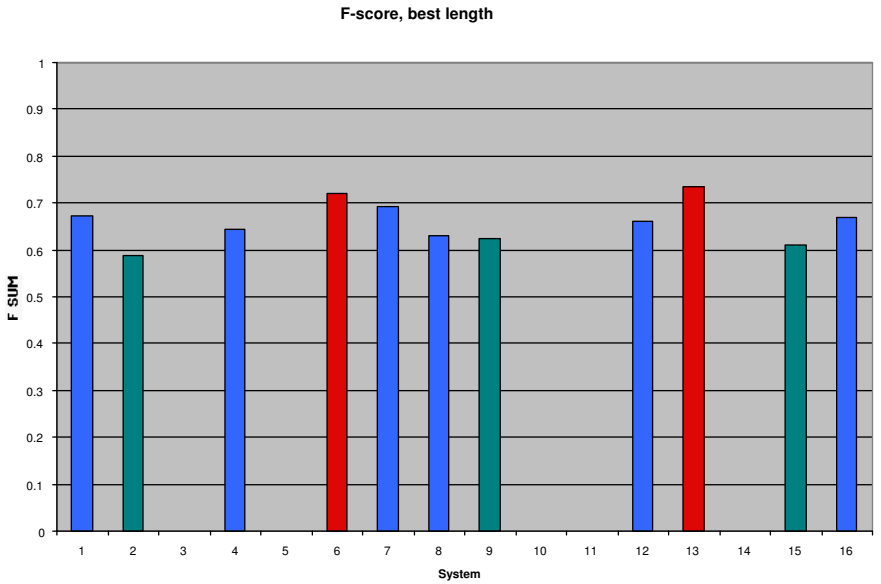
SUMMAC Ad Hoc (Query-Based) Test

- 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. Testers measure *R* and *P*: *how relevant are the summaries to their queries?*

     (many other measures as well)

- Results:

  3 levels of performance



Thanks !

Appendix 1
Sample Questions

Questions Answered by Slideset (1)

- **What dimensions (‚genres‘) are used to describe different kind of summaries?**
- **What are the "NLP/IE" and the "Statistics/IR" paradigms in summarization?**
  - What are the needs?
  - How do they relate to IR and IE?
  - What are the strengths, what the weaknesses of either one?
- **What extraction methods are there?**
- **Explain the contribution of lexical chains to summarization.**
- **What are cue phrases, how are they defined, and how are they used in summarization?**

Questions Answered by Slideset (2)

- What kinds of text interpretation are used for summarization?
- What are topic signatures, how are they defined, and how are they used in summarization?
- What difference would generation technology make to a summary?
- What measures are used to evaluate summarization systems?
- Evaluating summaries – when there are no previous summaries available – can be done according to different criteria. Define the measures of compression ration and retention ratio. Explain the "Q&A game" method and how retention is measured there.

Appendix 2
Corpora

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* having lengths from 161 to 725 words
- **Jing et al. (98)**
  - 40 newspaper articles from the TREC collection.

CORPORA IN SUMMARIZATION STUDIES (3)

- **For each text in each of the five corpora**
  - Human annotators determined the collection of salient sentences/clauses (Edmundson, Jing et al., Marcu) .
  - 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.**

Appendix 3
References

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.

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. 1998. Automated Text Summarization in SUMMARIST. In M. Maybury and I. Mani (eds), *Intelligent Scalable Summarization Text Summarization*. Forthcoming.

Hovy, E.H. 1998. Experiments in Evaluating Summarization. In prep.

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.

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.

Lenci, A., R. Bartolini, N. Calzolari, A. Auga, S. Busemann, E. Cartier, K. Chevrau, and J. Coch. 2002. Multilinual Summarization by Integrating Linguistic Resources into the MLIS-MUSI Project. In Proceedings of the 3rd International Conference on Language Ressources and Evaluation (LREC02). Las Palmas, Canary Islands, Spain.

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.

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.

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.

References (3)

Marcu, D. 1998. Improving Summarization Through Rhetorical Parsing Tuning. *Proceedings of the Workshop on Very Large Corpora*. Montreal, Canada.

Marcu, D. 1998. The Automatic Construction of Large-Scale Corpora for Summarization Research. In prep.

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

McKeown, K.R. and D.R. Radev. 1995. Generating Summaries of Multiple News Articles. In *Proceedings of the Eighteenth Annual International ACM Conference on Research and Development in Information Retrieval* (SIGIR), 74–82. Seattle, WA.

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.

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.

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.

References (4)

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.

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.

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.cs.columbia.edu/~radev/summarization/**
- **http://www1.cs.columbia.edu/~hjing/summarization.htm**