

# **Text Summarization**

based on a shortened  
tutorial presentation

by

Manabu Okumura

Precision & Intelligence Laboratory,

Tokyo Institute of Technology

# Table of Contents

1. What is text summarization?
2. Sentence extraction as a summarization method
3. Current trends in the summarization method
4. Evaluation of summaries

# Headline news — informing

The screenshot shows the TIME.com website interface. At the top left is the 'TIME.com' logo. To the right of the logo is a navigation bar with 'HOME' and 'SEARCH' links. Below the logo is a vertical menu with links to 'TIME Daily', 'News Wire', 'Editor's Letter', 'Comments', 'News Features', and 'Text Only'. Further down are links for 'Magazine', 'Community', and 'Special Reports'. A 'LIFE Picture of the Day' section is also visible. On the left side, there is a search box with 'Address' and 'Password' fields, and a 'go' button. Below the search box are two promotional banners: one for Microsoft Internet Explorer and another for AOL Instant Client. The main content area on the right features a date 'June 30, 1998' and a large headline 'U.S. Plane Fires a Missile On Iraq'. Below the headline is a sub-headline '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?' and a 'Full Story' link. To the right of the text is a photograph of an F-16 fighter jet in flight. Below the photo is a caption: 'Responding with Force: A U.S. Air Force F-16 flies over Kuwait. U.S. AIR FORCE/AP'. Below the main headline are three other news items: 'Starr Plays the Tripp Card' (with a sub-headline about a grand jury appearance), 'Down to Business in Shanghai' (with a sub-headline about President Clinton's visit), and 'Poll: Does the U.S. have the right to impose its idea of human rights on China?'. At the bottom of the main content area are two more items: '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?' and 'Boris Duels With the Duma' (with a sub-headline about Russian president Yeltsin).

**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  
AOL Instant Client

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.

From the tutorial of Hovy & Marcu in COLING/ACL'98

# Abstracts of papers — time saving

## An Incremental Interpreter for High-Level Programs with Sensing

Giuseppe De Giacomo

Dipartimento di Informatica e Sistemistica  
Università 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

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

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

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

to find a sequence with the right properties. This can involve considerable search when  $\delta$  is very nondeterministic, but much less search when  $\delta$  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.<sup>1</sup> Consider, for example, the following program:

From the tutorial of Hovy & Marcu in COLING/ACL'98

# Definition

Text summarization=  
the process to reduce the length or complexity  
of the original text, without losing the main  
content

# Input and Output

- Input
  - Source Text
  - Compression Rate or Summary Length
    - Rate = Summary Length/Source Length
- Output
  - Summary Text

# Current Applications

- Search Engines: summarize the information in hit lists retrieved by search engines
- Meeting Summarization: find out what happened at the conference I missed
- Hand-held devices: create a screen-sized summary of a book
- Aids for the Handicapped: compact the text and read it out for a blind
- ...

# Types 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-biased  
*...provides author's view vs. reflects user's interest.*
- Single-document vs. multi-document source  
*...based on one text vs. fuses together many texts.*

# Summary Function

- Indicative summaries  
provide a reference function for selecting documents for more in-depth reading
- Informative summaries  
cover all the important information in the source at some level of detail

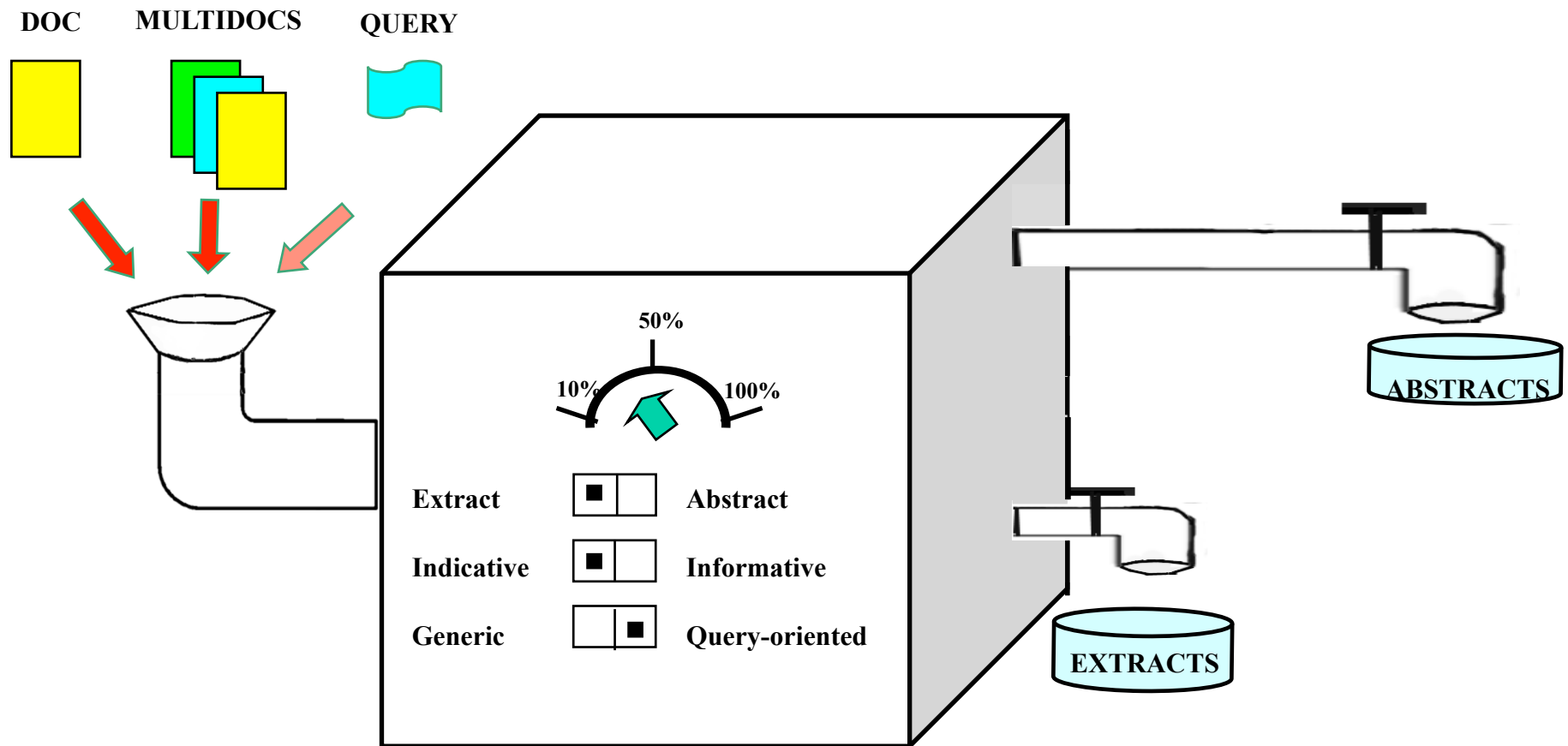
# Extract vs. Abstract

- An extract is a summary consisting entirely of material copied from the source
- An abstract is a summary at least some of whose material is not present in the source

# Aspects that describe summaries

- **Input** [Sparck Jones, 97]
  - *subject type*: domain
  - *genre*: newspaper articles, editorials, letters, reports...
  - *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**
  - *format*: paragraph, table, etc.
  - *style*: informative, indicative, ...

# A Summarization Machine



Modified from the tutorial of Hovy & Marcu in COLING/ACL'98

# Necessary NLP techniques

- Morphological analyzer  
(tokenizer, POS tagger)
- Parser (syntactic/rhetorical)
- Discourse interpretation  
(coreference resolution)
- Language generation
  - e.g., Controlled Natural Language

# Related Technologies

- Information Retrieval (IR)
  - Query-biased summarization
- Information Extraction (IE)
  - Key information is known beforehand (as a template)
- Question Answering (QA)
- Text Mining
- Text Classification
- Text Clustering

# Example of Information Extraction (Domain: Turnover)

(Input)

Sam Schwartz retired as executive vice president of the famous hot dog manufacturer, Hupplewhite Inc. He will be succeeded by Harry Himmelfarb.

→

(Output)

EVENT	leave job
PERSON	Sam Schwartz
POSITION	executive vice president
COMPANY	Hupplewhite Inc.
EVENT	start job
PERSON	Harry Himmelfarb
POSITION	executive vice president
COMPANY	Hupplewhite Inc.

# Example of QA

■ Q: What is the fastest car in the world?

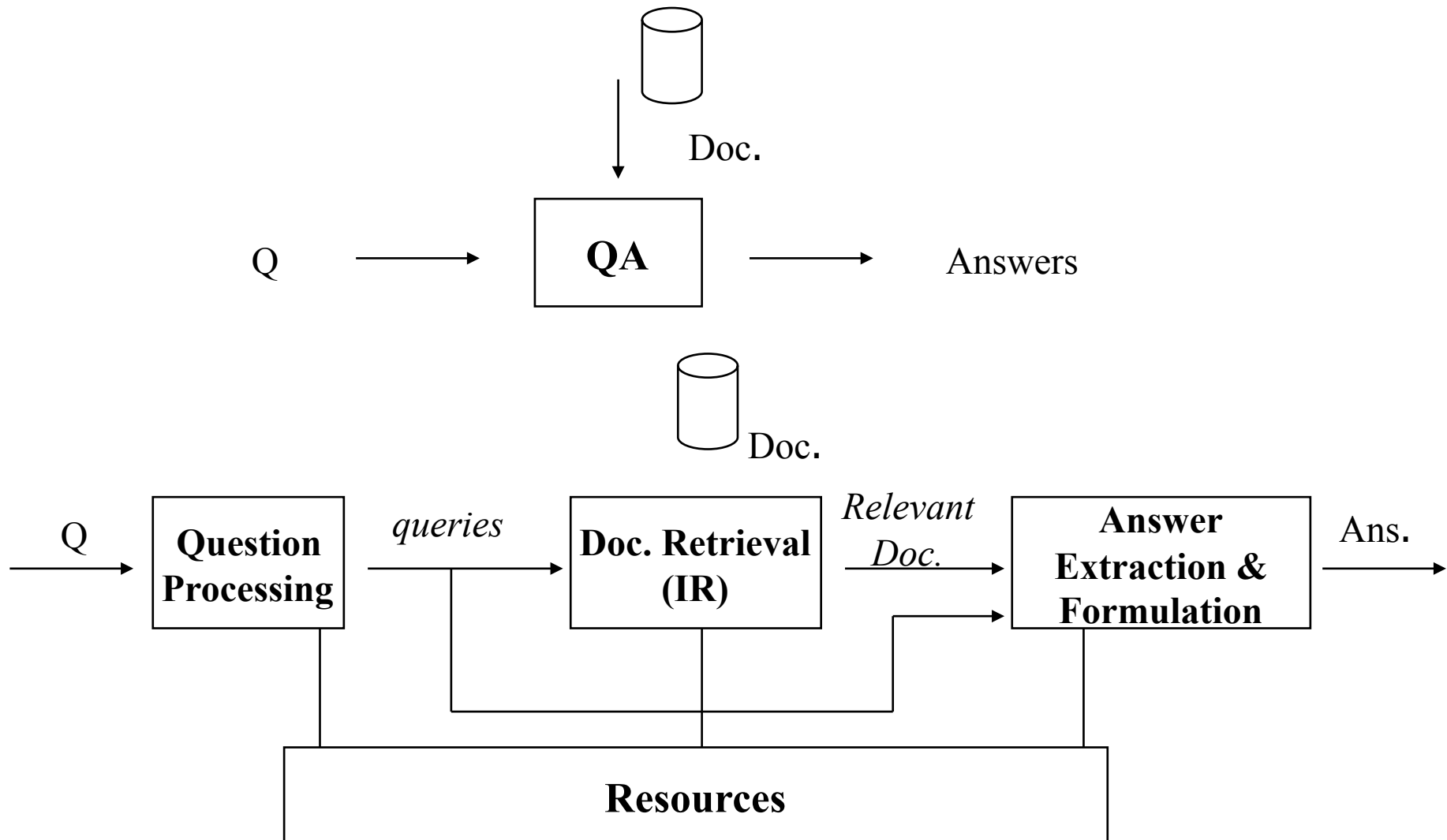
■ Correct answer:

.., the jaguar XJ220 is the dearest (Pounds 415,000), fastest (217 mph/ 350 kmh) and most sought-after car in the world.

■ Wrong answer:

.. will stretch Volkswagen's lead in the World's fastest-growing vehicle market. Demand is expected to soar in the next few years as more Chinese are able to afford their own cars.

# A Generic QA Architecture



# Text Summarization Methods (1/2)

Ideal method:

1. Understanding the text,
2. Transforming the analytical result of the text into the internal representation of a summary,  
(Finding the important parts of the analytical result)
3. Generating a summary from the internal representation

→

Current method:

1. Extracting the important parts (sentences) in the text,
2. Just arranging them in the order in the text

# Text Summarization Methods

(2/2)

- Sentence Extraction
- Sentence Simplification (Compression)
  - Try to shorten the text by removing unimportant parts in the sentences

# Information in text useful for sentence extraction

1. Word frequency in text,
2. Title of text,
3. Position of sentence,
4. Cue phrases in text,
5. Discourse structure of text,
6. Cohesion in text,
7. ...

[Paice,90]

# 1. Word frequency in text

- Important sentences contain content words that occur frequently in the text.
- Content words that frequently occur in text tend to indicate the topic of the text.
- The degree of the importance of words is calculated based on their frequency (tf).
- The degree of the importance of sentences is calculated based on the importance of words that they contain. [Luhn,58],[Zechner,96]

# Example2

1. 科学での言葉は、鋭利なメスのように物や事物の中に切り込み分けている。
2. 切り分けたうえで、またそれらの物を組みなおしてみる、それが科学での言葉の役割である。
3. だから科学の言葉は、科学者の自己を離れて物のほうに張りついている。
4. 科学が主観を排除して、客観的事実だけを示すというのは、このように科学では言葉が自己を離れているということによるのである。
5. これにたいして、詩での言葉は物を表現するときでも、自己を離れるということはない。
6. 物そのものが語っているように見えても、言葉は詩人の自己の手に握られているのである。
7. だからたとえその詩がどんなに客観的な描写に見えようとも、そこには詩人の自己の言葉が生きているのである。

石田春夫, 「学生のための自分学」より

言葉:7 科学:6 詩:4 自己:5

# Luhn's Summarization Method

- Set a limit  $L$  for the distance at which any two significant words could be considered as being significantly related
- Find out a portion in the sentence that is bracketed by significant words not more than  $L$  non-significant words apart
- Count the number of significant words contained in the portion. The result is significant factor related to  $S$



*Portion of sentence bracketed by and including significant words not more than four non-significant words apart. If eligible, the whole sentence is cited.*

# Genre-dependent text structure

Texts tend to have the structure that depends on their genre.

**Technical papers:** Introduction, main parts,  
Conclusion

**Newspapers:** headlines, subheads, body

## 2. Title of text

- Titles and headings are considered as the concise summaries of texts.
- Sentences that contain content words in titles and headings are important.

## 3. Position of sentence

### **Technical papers:**

- Topical sentences tend to occur at initial position of paragraphs.
- Important sentences tend to occur at beginning or end of text.

### **Newspapers:**

- It's a good method to extract the first some sentences in the body (Lead method).
- Lead method is said to be quite effective for newspaper summarization. More than 90% acceptability.[Brandow,95],[Wasson,98]

# Optimum Position Policy (OPP)

- **Claim:** Important sentences are located at positions that are genre-dependent; these positions can be determined automatically through training (Lin and Hovy, 97).
  - **Corpus:** 13000 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)

From the tutorial of Hovy & Marcu in COLING/ACL'98

# Opp (cont.)

– OPP for ZIFF corpus:

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

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

– OPP for *Wall Street Journal*:  $(T) > (P_1, S_1) > \dots$

## 4. Cue phrases in text

There are cue phrases that are positively/negatively correlated to important sentences.

**positive:** ‘In this paper, ‘In conclusion’, ‘our work’,... (in technical papers)

**negative:** conjunctives that indicate illustration, such as ‘for example’

## 5. Discourse structure of text

The discourse structure of a text can be constructed based on the surface information in the text, such as discourse markers (conjunctives), and the ‘centrality’ of the sentences in the structure reflects their importance.

[Miike,94],[Marcu,97]

# Rhetorical parsing

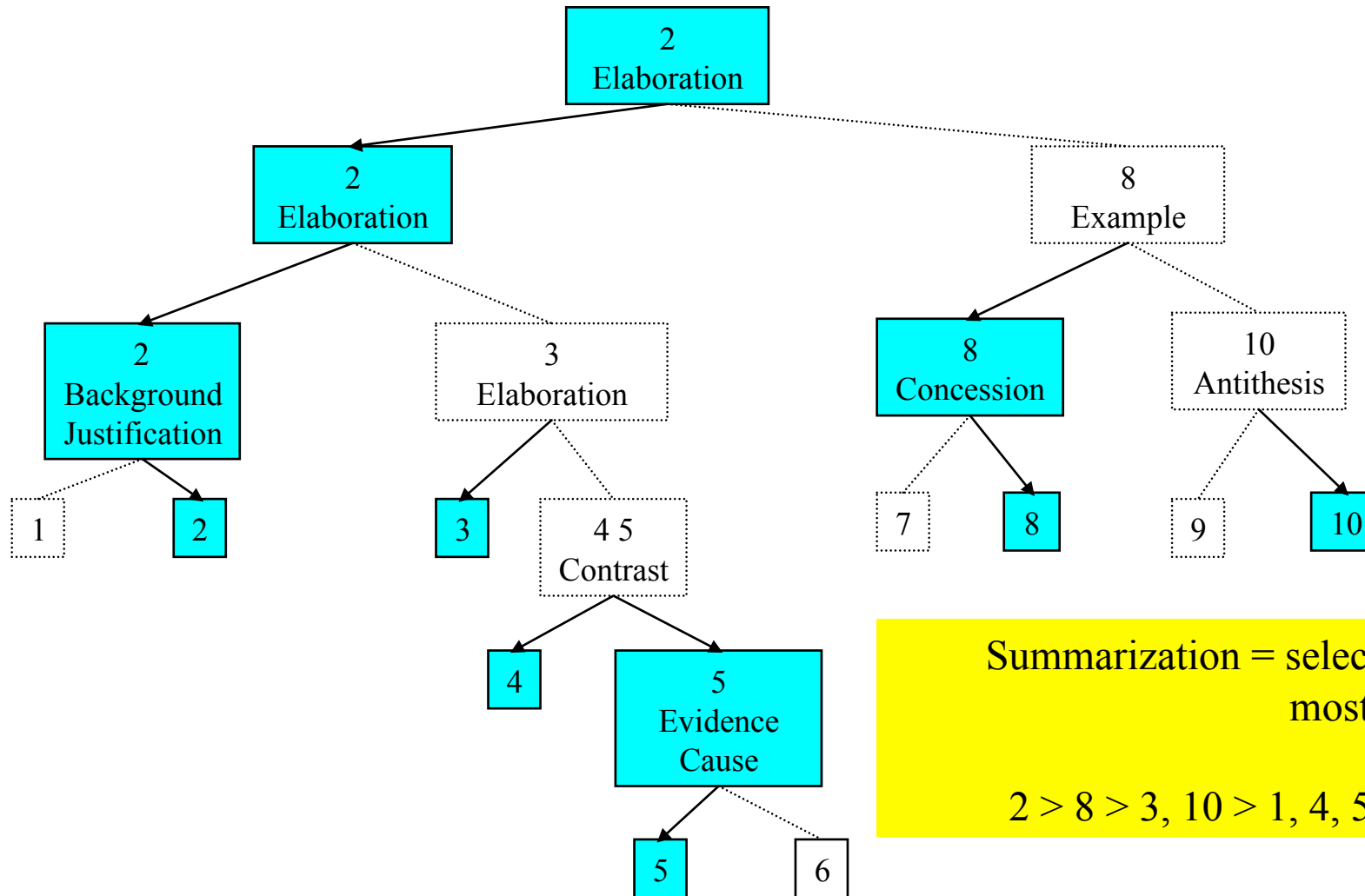
## (Marcu,97)

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

[*Although* the atmosphere holds a small amount of water, and water-ice clouds sometimes develop,<sup>7</sup>] [most Martian weather involves blowing dust or carbon dioxide.<sup>8</sup>] [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.<sup>9</sup>] [*Yet* even on the summer pole, {*where* the sun remains in the sky all day long,} temperatures never warm enough to melt frozen water.<sup>10</sup>]

From the tutorial of Hovy & Marcu in COLING/ACL'98

# Rhetorical parsing (2)



Summarization = selection of the most important units

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

From the tutorial of Hovy & Marcu in COLING/ACL'98

# Merits of the method

- Summaries of the various length can be obtained, at any depth of tree (structure).
- More coherent summary might be obtained, since summarization is based on the discourse structure.

## 6. Cohesion in text

Important sentences are the ones that are connected with many other sentences.

a) Use lexical cohesion to determine the connectedness between sentences

[Skorokhod'ko,72]

b) Use similarity between sentences to determine the connectedness between them

[Salton,96]

# Lexical cohesion

Semantic relationship between sentences is indicated by the use of related words in them

[Halliday & Hasan,76]

- #1# Apple Looking for a Partner
- #2# NEW YORK (Reuter) - Apple is actively looking for a friendly merger partner, according to several executives close to the company, the New York Times said on Thursday.
- #3# One executive who does business with Apple said Apple employees told him the company was again in talks with Sun Microsystems, the paper said.
- #4# On Wednesday, Saudi Arabia's Prince Alwaleed Bin Talal Bin Abdulaziz Al Saud said he owned more than five percent of the computer maker's stock, recently buying shares on the open market for a total of \$115 million.
- #5# Oracle Corp Chairman Larry Ellison confirmed on March 27 he had formed an independent investor group to gauge interest in taking over Apple.
- #6# The company was not immediately available to comment.

Table 1: Sample text with numbered text units

<i>Text units</i>		<i>Words shared</i>	<i>Score</i>	
#1#	#2#	Apple, look, partner	3	# 2
#1#	#3#	Apple, Apple	2	# 3
#1#	#4#		0	# 4
#1#	#5#	Apple	1	# 5
#1#	#6#		0	# 6
#2#	#3#	Apple, Apple, executive, company	4	
#2#	#4#		0	
#2#	#5#	Apple	1	
#2#	#6#	company	1	
#3#	#4#		0	
#3#	#5#	Apple, Apple	2	
#3#	#6#	company	1	
#4#	#5#		0	
#4#	#6#		0	
#5#	#6#		0	

Table 2: Measuring lexical cohesion in text unit pairs

# Sentence extraction by combining multiple information

How to combine multiple information?

The importance of each sentence is the weighted sum of the importance from multiple information (Linear combination).

How to weight each information?

- Human tuning [Edmundson,69]
- Automatic weighting using a set of summaries as training corpus
  - multiple regression analysis [Watanabe,96]

# New topics in text summarization

1. Abstraction vs. extraction  
Try to re-phrase content
2. Query-biased summarization vs. generic summarization  
Try to reflect the user's interest
3. Multi-document summarization vs. single-document summarization  
Try to fuse together the content of many texts
4. Sentence simplification vs. sentence extraction  
Try to shorten the text by removing unimportant parts in the sentences

# Summarization by Abstraction, Paraphrase

Abstract represents the content of the original text, by generalization or paraphrasing

Abstract generation=

extract (of the important concepts)+  
concept fusion +generation

# Concept fusion with conceptual hierarchy [Hovy,97]

John bought some vegetables, fruit, bread, and milk.

→

John bought some groceries.

# Query-biased vs. Generic

Summaries can be constructed solely from the content in the original text.

(static summary)



Summaries should be dynamic, reflecting the user's interest.

# Query-biased summaries for IR

- In case summaries are used for relevance judgment of texts in IR, they should reflect the query that the user inputs.
- More importance is given to the sentences that contain words in the query.

[Tombros,98]

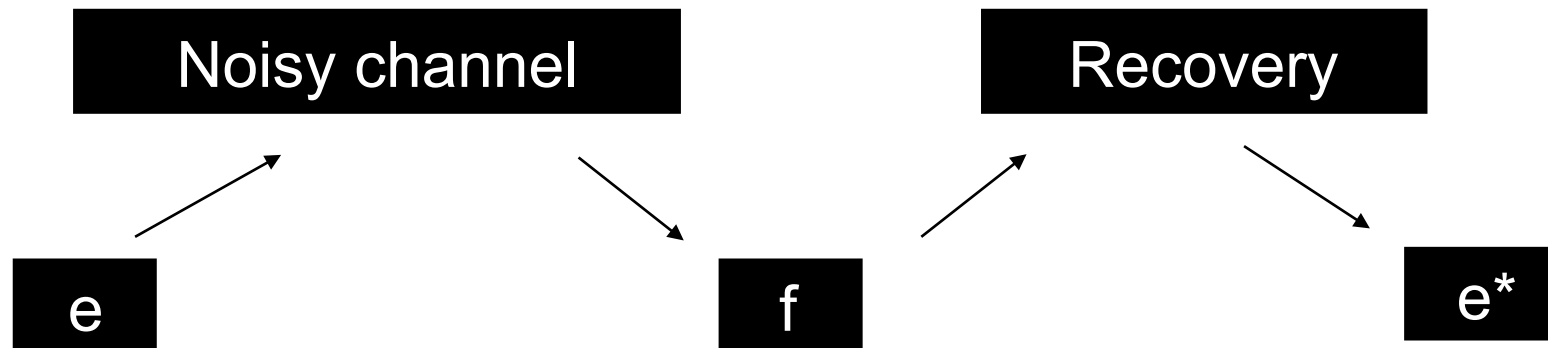
# Summarization using LM

- Source language: full document
- Target language: summary

From the tutorial of Radev in SIGIR'04

# Language modeling (1/2)

- Source/target language
- Coding process



- Noisy-Channel Model

# Language modeling(2/2)

$$e^* = \arg \max_e p(e | f) = \arg \max_e p(e).p(f | e)$$

$$p(E) = p(e_1).p(e_2 | e_1).p(e_3 | e_1e_2)...p(e_n | e_1...e_{n-1})$$

$$p(E) = p(e_1).p(e_2 | e_1).p(e_3 | e_2)...p(e_n | e_{n-1})$$

(Bigram model)

# [Berger & Mittal, 00]

- Gisting (OCELOT)

$$g^* = \arg \max_g p(g | d) = \arg \max_g p(g).p(d | g)$$

- content selection (preserve frequencies)
- word ordering (single words, consecutive positions)

# Framework of Text Summarization System

1. Sentence Extraction
  2. Duplicate Reduction
  3. Sentence Simplification
  4. Revision and/or Generation
- 2 is optional in single document summarization
  - 4 has not been fully studied

# Evaluation Types [Sparck Jones & Galliers, 96]

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

- compare against ideal output/source text
- criteria—*quality, informativeness*, etc.

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

- time to perform the task, accuracy of the task
- Reading Comprehension Tests [Morris et al., 92], IR, text categorization [SUMMAC, 98]

# Quality Evaluation

- Subjective grading
- 12 Quality Questions in DUC series
- Criteria: grammar, coherence, ...

# Examples of Quality Questions

1. About how many of the sentences are missing important components (e.g. the subject, main verb, direct object, modifier) - causing the sentence to be ungrammatical, unclear, or misleading?
2. About how many pronouns are there whose antecedents are incorrect, unclear, missing, or come only later?
3. About how many dangling conjunctions are there ("and", "however" ...)?

# Reading Comprehension Tests

- Human reads full texts or summaries, and then answers a test.
- The scores measuring the percentage of correct answers indicate the usefulness of summaries.
- If reading a summary allows a human to answer questions as accurately as he could with the source, the summary is highly informative.

# Resources (1)

## **Books:**

- Inderjeet Mani.  
Automatic Summarization, John Benjamins Publishing Company, 2001.
- Advances in Automatic Text Summarization, MIT Press, 1999.

## **Online bibliographies:**

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

# Resources (2)

## Online Tutorial Slides:

- Dragomir R. Radev.  
Text Summarization Tutorial, ACM SIGIR, 2004. <http://www.summarization.com/sigirtutorial2004.ppt>
- Eduary Hovy and Daniel Marcu.  
Automatic Text Summarization Tutorial, COLING/ACL, 1998.  
<http://www.isi.edu/~marcu/acl-tutorial.ppt>
- Horacio Saggion, Automatic text summarization: past, present, and future, IBERAMIA, 2004.  
<http://www.dcs.shef.ac.uk/~saggion/saggion04.PDF>

# Resources(3)

## **Multi-document Summarization System Softwares**

- <http://www.summarization.com/mead/>
- <http://www.clsp.jhu.edu/ws2001/groups/asmd/>