Automated Text Summarization

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

Based on slides by Ed Hovy and Daniel Marcu, USC-ISI

(! = may contain answers to exam questions)

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
June 30, 1998
U.S. Plane Fires a Missile On Iraq
An Iraqi fighter plane targets an Allied plane, and a U.S. F-16 responds quickly—with deadly force. Is this the showdown with Saddam on the way?

Star Wars the Tripp Card
The former president's grand jury appearance puts the squeeze on Mr. Lewinsky.

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

Pell: Dire the U.S. have the right to impose its own ideas of human rights on China?

Postcards from the Middle Kingdom: TMSL's Jay Stanwood says President Clinton's in full campaign mode in China. But the big question is, why isn't the president talking?

Fordy Deals With the Duma
If Russian president Yeltsin wants to make Russia a pole in the West, he should stop sending a feel 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 Comery 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

3:35am 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

Gianfranco De Giacomo
Department of Informatics and Mathematics
Università di Roma "La Sapienza"
Via Salaria 113, 00198 Rome, Italy
dgiacomo@dis.uniroma1.it

Hector Levesque
Department of Computer Science
University of Toronto
Toronto, Canada M5S 3G3
levesque@cs.toronto.edu

Abstract

Like classical planning, the execution of high-level planner programs requires a sequence of steps to be a final
solution. While, in a classical planner, this sequence is determined at design time, in the case of
the high-level programs, this sequence is determined at runtime. Moreover, our approach requires
that the high-level planning be an integral part of the final solution. We have developed a system for
the execution of high-level planning programs in which the high-level programs can be executed in an
incremental manner. The high-level programs are translated into a form that is suitable for execution
by a classical planner. The incremental interpreter is used to find a sequence of actions that is
consistent with the requirements of the high-level program. The interpreter is also used to
monitor the progress of the high-level program and to provide feedback to the user.

Introduction

In [4], we described a system for providing high-level
planning in a classical planner. We showed how the use of
high-level planning can improve the efficiency of classical
planning and to reduce the amount of planning that needs to be
performed.

Assumption 1: Legal( π(θ, θ)) ∧ δ(θ, θ)

where 0 is the goal being planned for, we look for a
sequence of actions such that

Assumption 2: δ(θ, θ), 0 (θ, θ)

to find a sequence with the right properties. This can be
achieved by using a high-level planner to generate a
sequence of actions that satisfies the requirements of the
high-level program. The high-level program is then
translated into a form that is suitable for execution by a
classical planner. The classical planner is then used to
execute the high-level program in an incremental manner.

Graphical Maps — Orienting

Route:

Origin: Del Mar, CA

Destination: Marina del Rey, CA

© 1999 GeoSystems Global Corp., © 1998 NavTech

© 1999 GeoSystems Global Corp., © 1998 NavTech

Author: Stephan Busemann

Language Technology I, WS 2011/2012, 5

Author: Stephan Busemann

Language Technology I, WS 2011/2012, 6
Textual Directions — Planning

Door to Door Directions:

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

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

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 rule-based 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, don’t spend more than 50 words each

**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: children’s story**

**reader3:** your 4-year-old niece.

**reader4:** amazon customer.

---

**90 Soldiers Arrested After Coup Attempt In Tribal Homeland**

MMAABATHO, 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 Motlhabe 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.
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*...
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

<table>
<thead>
<tr>
<th>NLP/IE:</th>
<th>IR/Statistics:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Approach:</strong> try to ‘understand’ text—re-represent content using ‘deeper’ notation; then manipulate that.</td>
<td><strong>Approach:</strong> operate at lexical level—use word frequency, collocation counts, etc.</td>
</tr>
<tr>
<td><strong>Need:</strong> rules for text analysis and manipulation, at all levels.</td>
<td><strong>Need:</strong> large amounts of text.</td>
</tr>
<tr>
<td><strong>Strengths:</strong> higher quality; supports abstracting.</td>
<td><strong>Strengths:</strong> robust; good for query-oriented summaries.</td>
</tr>
<tr>
<td><strong>Weaknesses:</strong> speed; still needs to scale up to robust open-domain summarization.</td>
<td><strong>Weaknesses:</strong> lower quality; inability to manipulate information at abstract levels.</td>
</tr>
</tbody>
</table>

## 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).

**Mrs. Coolidge:** “What did the preacher preach about?”
**Coolidge:** “Sin.”
**Mrs. Coolidge:** “What did he say?”
**Coolidge:** “He’s against it.”
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
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**

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)

<table>
<thead>
<tr>
<th>Individual contribution</th>
<th>Cumulative contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>• (Edmundson, 68)</td>
<td>• (Edmundson, 68)</td>
</tr>
<tr>
<td>– 52% recall &amp; precision in combination with title (25% lead baseline)</td>
<td>– the best individual method</td>
</tr>
<tr>
<td>• (Kupiec et al., 95)</td>
<td>• Kupiec et al., 95</td>
</tr>
<tr>
<td>– 33% recall &amp; precision</td>
<td>– the best individual method</td>
</tr>
<tr>
<td>– (24% lead baseline)</td>
<td></td>
</tr>
<tr>
<td>• (Teufel and Moens, 97)</td>
<td>• (Teufel and Moens, 97)</td>
</tr>
<tr>
<td>– 32% recall and precision (28% lead baseline)</td>
<td>– increased performance by 10% when combined with the cue-based method</td>
</tr>
</tbody>
</table>

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:

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

\((T=\text{title}; \, P=\text{paragraph}; \, S=\text{sentence})\)

– OPP for Wall Street Journal: \((T) > (P_1, S_1) > \ldots\)

– **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)

<table>
<thead>
<tr>
<th>Individual contribution</th>
<th>Cumulative contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Edmundson, 68)</td>
<td>(Edmundson, 68)</td>
</tr>
<tr>
<td>– 40% recall &amp; precision</td>
<td>(– increased performance by 8% when combined with the title- and cue-based methods.)</td>
</tr>
<tr>
<td>(Teufel and Moens, 97)</td>
<td>(Teufel and Moens, 97)</td>
</tr>
<tr>
<td>– 21.7% recall &amp; precision</td>
<td>(– increased performance by 3% when combined with cue-, location-, position-, and word-frequency-based methods.)</td>
</tr>
</tbody>
</table>

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)

- **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.

(Luhn, 58)
Word-Frequency-Based Method (2)

<table>
<thead>
<tr>
<th>Individual contribution</th>
<th>Cumulative contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>• (Edmundson, 68)</td>
<td>• (Edmundson, 68)</td>
</tr>
<tr>
<td>– 36% recall &amp; precision</td>
<td>– decreased performance by 7%</td>
</tr>
<tr>
<td>(25% lead baseline)</td>
<td>when combined with other methods</td>
</tr>
<tr>
<td>• (Kupiec et al., 95)</td>
<td>• (Kupiec et al., 95)</td>
</tr>
<tr>
<td>– 20% recall &amp; precision</td>
<td>– decreased performance by 2%</td>
</tr>
<tr>
<td>(24% lead baseline)</td>
<td>when combined...</td>
</tr>
<tr>
<td>• (Teufel and Moens, 97)</td>
<td>• (Teufel and Moens, 97)</td>
</tr>
<tr>
<td>– 17% recall &amp; precision</td>
<td>– increased performance by 0.2%</td>
</tr>
<tr>
<td>(28% lead baseline)</td>
<td>when combined...</td>
</tr>
</tbody>
</table>

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

- 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.

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.

- Computing LCs efficiently (Silber and McCoy, 02)
  - Using WordNet synsets and relations
  - Evaluation design by using a Text/Summary corpus
    - Most strong chains should be in the summary
    - Each noun in the summary should be used in the same sense as some word in a strong chain in the original document

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)
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, [Mars experiences frigid weather conditions.]* [Surface temperatures typically average about –60 degrees Celsius (–76 degrees Fahrenheit) at the equator and can dip to –123 degrees C near the poles.]* [Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion.]* [but any liquid water formed that way would evaporate almost instantly]* [because of the low atmospheric pressure.]

*Although* the atmosphere holds a small amount of water, and water-ice clouds sometimes develop, [most Martian weather involves blowing dust or carbon dioxide.]* [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.]* [Yet even on the summer pole, where the sun remains in the sky all day long,] temperatures never warm enough to melt frozen water.*
Rhetorical Parsing (2)

- Use discourse markers to hypothesize rhetorical relations
  - \texttt{rhet\_rel(CONTRAST, 4, 5)} \otimes \texttt{rhet\_rel(CONTRAST, 4, 6)}
  - \texttt{rhet\_rel(EXAMPLE, 9, [7,8])} \otimes \texttt{rhet\_rel(EXAMPLE, 10, [7,8])}

- Use semantic similarity to hypothesize rhetorical relations
  - if similar($u_1, u_2$) then
    \texttt{rhet\_rel(ELABORATION, $u_2$, $u_1$)} \otimes \texttt{rhet\_rel(BACKGROUND, $u_1$, $u_2$)}
  - else
    \texttt{rhet\_rel(JOIN, $u_1$, $u_2$)}
  - \texttt{rhet\_rel(JOIN, 3, [1,2])} \otimes \texttt{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
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**:
  - (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**:

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MESSAGE:ID</td>
<td>TSL-COL-0001</td>
</tr>
<tr>
<td>SECSOURCE:SOURCE</td>
<td>Reuters</td>
</tr>
<tr>
<td>SECSOURCE:DATE</td>
<td>26 Feb 93</td>
</tr>
<tr>
<td>INCIDENT:DATE</td>
<td>26 Feb 93</td>
</tr>
<tr>
<td>INCIDENT:LOCATION</td>
<td>World Trade Center</td>
</tr>
<tr>
<td>INCIDENT:TYPE</td>
<td>Bombing</td>
</tr>
<tr>
<td>HUM TGT:NUMBER</td>
<td>AT LEAST 5</td>
</tr>
</tbody>
</table>
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.
   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:
    
    \[ \text{extract\_length} = 2.76 \times \text{abstract\_length} \]
Some Types of Interpretation

- **Concept generalization:**
  Sue ate apples, pears, and bananas ⇒ Sue ate fruit

- **Mereonymy 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.

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):

  
  \[
  \text{signature} = \{\text{head} (c_1, f_1) (c_2, f_2) \ldots\}
  \]

  \[
  \text{restaurant} \Leftarrow \text{waiter + menu + food + eat...}
  \]

- (inverse of query expansion in IR.)
Example Signatures

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

Overview

1. Motivation.
2. Genres and types of summaries.
3. Approaches and paradigms.
   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.

Multiple sources and disagreement

Explicit mentioning of “no information”.

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

Overview

1. Motivation.
2. Genres and types of summaries.
3. Approaches and paradigms.
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: \( CR = \frac{\text{length } S}{\text{length } T} \)
  Retention Ratio: \( RR = \frac{\text{info in } S}{\text{info in } T} \)

• 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):**

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>approx. 10</td>
<td>approx. 150</td>
<td>approx. 1100</td>
</tr>
</tbody>
</table>

- Depends on the subject’s knowledge of the topic
Q&A Evaluation

- **Can you focus on the important stuff?**
  The Q&A Game—can be tailored to your interests!

- **Measure core information 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 topic.
  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...

![Graph showing normalized answer recall](image-url)
Thanks!

Appendix 1
Sample Questions
Questions Answered by Slideset (1)

• What dimensions (‘genres’) are used to describe different kinds 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 ratio 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.
## References (1)


## References (2)


References (3)


References (4)


