NLP

Introduction to NLP

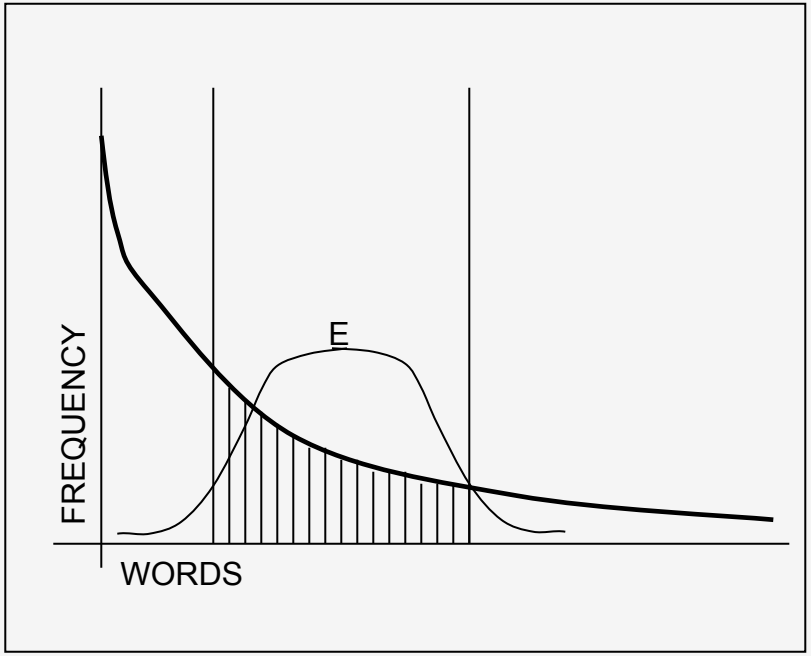
Summarization Techniques 1/3

Baxendale (1958)

- Positional method
  - Analysis of 200 paragraphs
  - Pick the first and last sentences of the paragraph
  - That's where the topic sentences are located
  - Naïve but decent approach

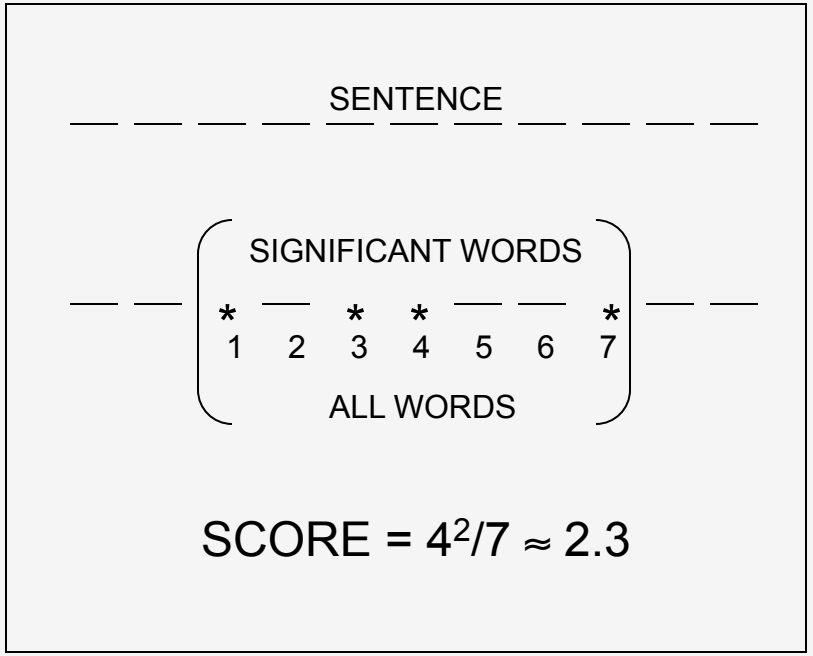
Luhn (1958)

- Technical documents
- Stemming
- Stop words are removed
- Frequency of content terms



Luhn (1958)

- Sentence-level significance factor
- Look for concentrations of salient content terms



Edmundson (1969)

- Technical documents
- Position and frequency
- Cue words (bonus and stigma words)
  - Significant, hardly, impossible
- Document structure
  - Is the sentence a title or heading or right under one of these
- Linear combination of the four features

$$\alpha_1 C + \alpha_2 K + \alpha_3 T + \alpha_4 L$$

Frump (deJong 1979, 1982)

- Knowledge-based
- Slot-filling based on UPI news stories
- Based on 50 sketchy scripts
- Inputs matched to scripts based on manually selected keywords
- Difficult to port to other domains
- Missing scripts for many inputs

Frump

$demonstration script

- The demonstrators arrive at the demonstration location.
- The demonstrators march.
- Police arrive on the scene.
- The demonstrators communicate with the target of the demonstration.
- The demonstrators attack the target of the demonstration.
- The demonstrators attack the police.

G. DeJong (1979) FRUMP: Fast Reading Understanding and Memory Program

Paice (1990)

- Survey up to 1990
- Techniques that (mostly) failed
  - Syntactic criteria (Earl 1970)
  - Indicator phrases
- Problems with extracts
  - Lack of balance
  - Lack of cohesion

Paice (1990)

- Lack of balance
  - later approaches based on text rhetorical structure
- Lack of cohesion
  - anaphoric reference
  - lexical or definite reference
  - rhetorical connectives
  - recognition of anaphors [Liddy et al. 87]
  - Example: “that” is
    - *nonanaphoric* if preceded by a research-verb (e.g., “demonstrat-”),
    - *nonanaphoric* if followed by a pronoun, article, quantifier,…,
    - *external* if no later than 10th word, else
    - *internal*

Brandow et al. (1995)

- ANES: commercial news from 41 publications
- "Lead" achieves acceptability of 90% vs. 74.4% for "intelligent" summaries
- 20,997 documents
- words selected based on tf*idf (term frequency * inverse document frequency)
- sentence-based features:
  - signature words
  - location
  - anaphora words
  - length of abstract
- Sentences with no signature words are included if between two selected sentences
- Evaluation done at 60, 150, and 250 word length
- Non-task-driven evaluation:
  - "Most summaries judged less-than-perfect would not be detectable as such to a user"

Kupiec et al. (1995)

- First trainable method
  - 20% extract
  - 188 documents from scientific journals
  - Naïve Bayes classifier
- New features
  - Sentence length (|S|>5)
  - Presence of uppercase words (except common acronyms)
  - Thematic words
  - Set of 26 manually fixed phrases
  - Sentence position in paragraph

Kupiec et al. (1995)

- Uses Naïve Bayesian classifier

$$P(s \in S \mid F_1, F_2, \ldots F_k) = \frac{P(F_1, F_2, \ldots F_k \mid s \in S) P(s \in S)}{P(F_1, F_2, \ldots F_k)}$$

- Assuming statistical independence

$$P(s \in S \mid F_1, F_2, \ldots F_k) = \frac{\prod_{j=1}^{k} P(F_j \mid s \in S) P(s \in S)}{\prod_{j=1}^{k} P(F_j)}$$

Kupiec et al. (1995)

- Performance:
  - For 25% summaries, 84% precision
  - For smaller summaries, 74% improvement over Lead

Summons (McKeown & Radev 1995)

- First work on multi-document summarization
- Approach
  - Knowledge-based
  - Information extraction (MUC templates)
  - Text generation

Summons (McKeown & Radev 1995)

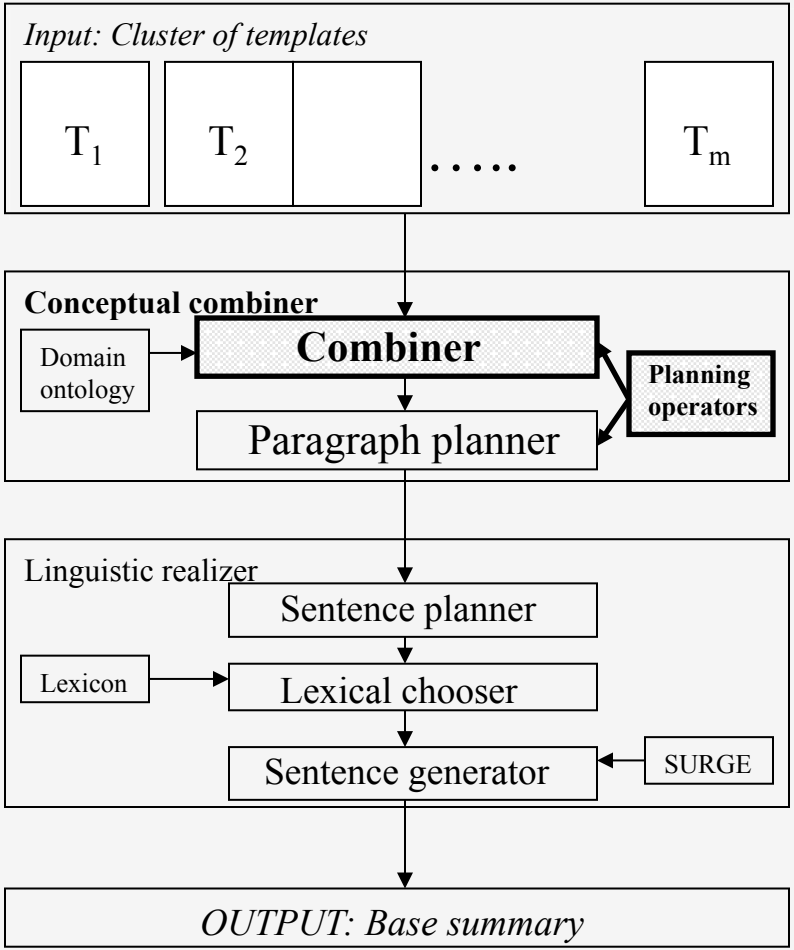
NICOSIA, Cyprus (AP) – Two bombs exploded near government ministries in Baghdad, but there was no immediate word of any casualties, Iraqi dissidents reported Friday. There was no independent confirmation of the claims by the Iraqi National Congress. Iraq's state-controlled media have not mentioned any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

Multiple sources and disagreement

Explicit mentioning of "no information".

Summons

MESSAGE: ID	TST3-MUC4-0010
MESSAGE: TEMPLATE	2
INCIDENT: DATE	30 OCT 89
INCIDENT: LOCATION	EL SALVADOR
INCIDENT: TYPE	ATTACK
INCIDENT: STAGE OF EXECUTION	ACCOMPLISHED
INCIDENT: INSTRUMENT ID	
INCIDENT: INSTRUMENT TYPE	
PERP: INCIDENT CATEGORY	TERRORIST ACT
PERP: INDIVIDUAL ID	"TERRORIST"
PERP: ORGANIZATION ID	"THE FMLN"
PERP: ORG. CONFIDENCE	REPORTED: "THE FMLN"
PHYS TGT: ID	
PHYS TGT: TYPE	
PHYS TGT: NUMBER	
PHYS TGT: FOREIGN NATION	
PHYS TGT: EFFECT OF INCIDENT	
PHYS TGT: TOTAL NUMBER	
HUM TGT: NAME	
HUM TGT: DESCRIPTION	"1 CIVILIAN"
HUM TGT: TYPE	CIVILIAN: "1 CIVILIAN"
HUM TGT: NUMBER	1: "1 CIVILIAN"
HUM TGT: FOREIGN NATION	
HUM TGT: EFFECT OF INCIDENT	DEATH: "1 CIVILIAN"
HUM TGT: TOTAL NUMBER	


Input: Cluster of templates
T1
T2
.....
Tm
Conceptual combiner
Domain ontology
Combiner
Planning operators
Paragraph planner
Linguistic realizer
Sentence planner
Lexicon
Lexical chooser
Sentence generator
SURGE
OUTPUT: Base summary

Summons

MESSAGE: ID	TST-REU-0001		MESSAGE: ID	TST-REU-0002
SECSOURCE: SOURCE		Reuters	SECSOURCE: SOURCE	Reuters
SECSOURCE: DATE		**March 3, 1996 11:30**	SECSOURCE: DATE	**March 4, 1996 07:20**
PRIMSOURCE: SOURCE			PRIMSOURCE: SOURCE	**Israel Radio**
INCIDENT: DATE		**March 3, 1996**	INCIDENT: DATE	**March 4, 1996**
INCIDENT: LOCATION		**Jerusalem**	INCIDENT: LOCATION	**Tel Aviv**
INCIDENT: TYPE		Bombing	INCIDENT: TYPE	Bombing
HUM TGT: NUMBER		**"killed: 18"**	HUM TGT: NUMBER	**"killed: at least 10"**
		"wounded: 10"		**"wounded: more than 100"**
PERP: ORGANIZATION ID			PERP: ORGANIZATION ID	
MESSAGE: ID	TST-REU-0003		MESSAGE: ID	TST-REU-0004
SECSOURCE: SOURCE		Reuters	SECSOURCE: SOURCE	Reuters
SECSOURCE: DATE		**March 4, 1996 14:20**	SECSOURCE: DATE	**March 4, 1996 14:30**
PRIMSOURCE: SOURCE			PRIMSOURCE: SOURCE	
INCIDENT: DATE		**March 4, 1996**	INCIDENT: DATE	**March 4, 1996**
INCIDENT: LOCATION		**Tel Aviv**	INCIDENT: LOCATION	**Tel Aviv**
INCIDENT: TYPE		Bombing	INCIDENT: TYPE	Bombing
HUM TGT: NUMBER		**"killed: at least 13"**	HUM TGT: NUMBER	**"killed: at least 12"**
		"wounded: more than 100"		**"wounded: 105"**
PERP: ORGANIZATION ID		**"Hamas"**	PERP: ORGANIZATION ID	

Summons

Reuters reported that 18 people were killed on *Sunday* in a bombing in Jerusalem. *The next day*, a bomb in Tel Aviv killed at least 10 people and wounded 30 according to Israel radio. Reuters reported that *at least 12 people* were killed and *105* wounded *in the second incident*. *Later the same day*, Reuters reported that Hamas has claimed responsibility for the act.

Summons

- If there are two templates
  AND
  the location is the same
  AND
  the time of the second template is after the time of the first template
  AND
  the source of the first template is different from the source of the second template
  AND
  at least one slot differs
  THEN
  combine the templates using the contradiction operator...

Summons

Change of perspective

Precondition:
The same source reports a change in a small number of slots

> March 4th, Reuters reported that a bomb in Tel Aviv killed at least 10 people and wounded 30. *Later the same day*, Reuters reported that *exactly 12 people* were *actually* killed and *105* wounded.

Summons

Contradiction

Precondition:
Different sources report contradictory values for a small number of slots

> The afternoon of February 26, 1993, Reuters reported that a suspected bomb killed *at least six people* in the World Trade Center. *However*, Associated Press announced that *exactly five people* were killed in the blast.

Summons

Refinement

On Monday morning, Reuters announced that a suicide bomber killed at least 10 people in Tel Aviv. *In the afternoon*, Reuters reported that *Hamas* claimed responsibility for the act.

Agreement

The morning of March 1st 1994, *both* UPI and Reuters reported that a man was kidnapped in the Bronx.

Summons

Generalization

According to UPI, three terrorists were arrested in Medellín last Tuesday. Reuters announced that the police arrested two drug traffickers in Bogotá the next day.

A total of five criminals were arrested in Colombia last week.

NLP