Automatic Summarization

ANLP
Horacio Rodríguez
Outline

• **Introduction**
• Problems of AS
• Basic Architectures
• Evaluations
• Single-document Systems
• Multi-document Systems
• Headline extraction
• Query-based Systems
• Multimedia and Multilingual Systems
• Multitask Systems
A summary is a reductive transformation of a source text into a summary text by extraction or generation

- Sparck-Jones, 2001
Introduction 2

- A possible objective

Illustration of the power of human abstracts

Mrs. Coolidge: What did the preacher discuss in his sermon?
President Coolidge: Sin.
Mrs. Coolidge: What did he say?
President Coolidge: He said he was against it.

- Bartlett’s Quotations (via Graeme Hirst)
Introduction

- Locate the fragments of a text relevant (in general or given the user needs) and produce a summary of them

- Sum vs IE
  - IE
    - Structure to be extracted is predefined
      - “I know what I want, look for it”
  - Sum
    - A previous definition of what we want is not (always) provided
      - “What is interesting here”
What to read

- **Tutorials**
  - E.Hovy, D. Marcu (Coling-ACL, 1998)
  - Maybury, Mani (EACL-ACL, 2001)
  - Mani (RANLP, 2003)
  - Lin (ACL, 2004)
  - Saggion (LREC, 2006)
  - [Lloret, Palomar, 2012]
  - [Saggion, Poibeau, 2013]

- **Surveys**

- **Books**
  - I. Mani, M. Maybury (1999) (eds), Mani (2001)
What to read

• Thesis
  - Barzilay (master, 1997), Barzilay (2003), C-Y. Lin (1997),
  - Tucker (1999), Kan (2003), Alonso (2005),
  - M. Fuentes (2008), E. Lloret (2011)

• Challenges
  - TIPSTER (http://www.tipster.org/index.htm)
  - SUMMAC (http://www.tipster.org/summcall.htm) 1998
  - Multilingual Summarization Evaluation (MSE) 2005, 2006
  - TAC (http://www.nist.gov/tac/2008/summarization/)

• Conferences and workshops
  - ANLP/NAACL summarization workshop 2000
  - NAACL summarization workshop 2001
  - Crossing barriers, RANLP 2005
Material

- U. Ottawa: The Text Summarization Project
  - http://www.site.uottawa.ca/tanka/ts.html
  - Ken Barker, Terry Copeck, Peter Turney, Stan Szpakowicz

- U. Columbia
  - Kathy McKeown, Judith Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, Hongyan Jing, Min-Yen Kan, Michael Elhadad, Ani Nenkova, Barry Schiffman, Smaranda Muresan

- U. Michigan
  - Dragomir Radev

- IDA, U. Maryland
  - Conroy, Schlesinger, O'Leary

- UPF
  - Horacio Saggion
Material 2

- Information Sciences Institute of the University of Southern California (USC/ISI)
  - http://www.isi.edu/natural-language/
  - Eduard Hovy, Kevin Knight, Chin-Yew Lin, Daniel Marcu, Hal Daumé III
- CMU (J. Carbonell, J. Goldstein)
- Other groups & people
<table>
<thead>
<tr>
<th>System Group</th>
<th>HE DUC02,03,04</th>
<th>MDS DUC02,03,04</th>
<th>Crosslingual DUC03, MSE05,06</th>
<th>qMDS DUC04,05,06</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLASSY IDA</td>
<td></td>
<td>[SOC+02]</td>
<td>[CSSO04]</td>
<td>[CSSO04]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[DCS+03]</td>
<td>[CGO05]</td>
<td>[CSG05]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[CSSO04]</td>
<td>[COS06]</td>
<td>[CSOe06]</td>
</tr>
<tr>
<td>DEMS-MG</td>
<td>[MBBG+02]</td>
<td>[BGEH+04]</td>
<td>[BGEH+04]</td>
<td>[BGEH+04]</td>
</tr>
<tr>
<td>Columbia</td>
<td>[NSS+03]</td>
<td>[BGEH+04]</td>
<td>[SE05]</td>
<td>[BG05]</td>
</tr>
<tr>
<td></td>
<td>[BGEH+04]</td>
<td></td>
<td></td>
<td>[SBG06]</td>
</tr>
<tr>
<td>GISTexter LCC's</td>
<td>[HL02]</td>
<td>[HL02]</td>
<td>[LHHN04]</td>
<td>[LPH03]</td>
</tr>
<tr>
<td></td>
<td>[LPH03]</td>
<td>[LPH03]</td>
<td>[LHHN04]</td>
<td>[LHAH05]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[LHHN04]</td>
<td></td>
<td>[LHR+06]</td>
</tr>
<tr>
<td>GLEANS ISI</td>
<td>[MDE+02]</td>
<td>[MDE+02]</td>
<td></td>
<td>[DM05]</td>
</tr>
<tr>
<td>(Marcu, Daume III)</td>
<td>[DM04]</td>
<td>[DM04]</td>
<td>[DM06]</td>
<td>[DM5]</td>
</tr>
<tr>
<td>KULeuven</td>
<td>[ABM02]</td>
<td>[ABM02]</td>
<td></td>
<td>[AMJM04]</td>
</tr>
<tr>
<td></td>
<td>[AMDB03]</td>
<td>[AMDB03]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[AMJM04]</td>
<td>[AMJM04]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEAD UMichigan</td>
<td>[OWR02]</td>
<td>[OWR02]</td>
<td>[ER04]</td>
<td>[ER04]</td>
</tr>
<tr>
<td></td>
<td>[ROQ0T03]</td>
<td>[ROQ0T03]</td>
<td>[ER04]</td>
<td></td>
</tr>
<tr>
<td>NeATS ISI</td>
<td>[LH02]</td>
<td>[LH02]</td>
<td></td>
<td>[ZLH05]</td>
</tr>
<tr>
<td>(Hovy, Lin, Zhou)</td>
<td>[ZH03]</td>
<td>[ZH03]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OGI/OHSU</td>
<td>[RF05]</td>
<td>[FHYH05]</td>
<td>[FR06]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[FL03]</td>
<td>[SG04]</td>
<td>[FL03]</td>
<td>[Sag05a]</td>
</tr>
<tr>
<td></td>
<td>[SG04]</td>
<td>[Sag05a]</td>
<td>[FL03]</td>
<td>[Sag05b]</td>
</tr>
<tr>
<td>SumUM USheffield</td>
<td>[AR03]</td>
<td>[AMSG04]</td>
<td>[DCW06]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[TAJMGM05]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[AOMSo06]</td>
</tr>
</tbody>
</table>

Table 2.1: Systems or Groups working in several tasks
<table>
<thead>
<tr>
<th>System</th>
<th>Languages</th>
<th>Type</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centrifier</td>
<td>English, French, German</td>
<td>MDS</td>
<td><a href="http://centrifier.cs.columbia.edu/centrifier.cgi">http://centrifier.cs.columbia.edu/centrifier.cgi</a></td>
</tr>
<tr>
<td>Island InText</td>
<td>English</td>
<td>SDS</td>
<td><a href="http://www.islandsoft.com/orderform.html">http://www.islandsoft.com/orderform.html</a></td>
</tr>
<tr>
<td>Kmartime</td>
<td>Korean</td>
<td>English</td>
<td>[<a href="http://publah.kmartime.ac.kr/demo/">http://publah.kmartime.ac.kr/demo/</a>] (<a href="http://publah.kmartime.ac.kr/demo/">http://publah.kmartime.ac.kr/demo/</a>)</td>
</tr>
<tr>
<td>NewsInEssence</td>
<td>English, Chinese</td>
<td>MDS</td>
<td><a href="http://www.newsinessences.com/nie.cgi">http://www.newsinessences.com/nie.cgi</a></td>
</tr>
<tr>
<td>MS-Word AutoSummarize</td>
<td>supposedly any language</td>
<td>SDS</td>
<td>included in MS-Word</td>
</tr>
<tr>
<td>MS-Word AutoSummarizer</td>
<td>English, French, Spanish, German, Italian, Portuguese, Japanese, Chinese, Korean, Arabic, Greek, Dutch, Norweigan and Russian</td>
<td>SDS</td>
<td><a href="http://www.pertinence.net">http://www.pertinence.net</a></td>
</tr>
<tr>
<td>Summ-It</td>
<td>probably English only</td>
<td>pasted text</td>
<td><a href="http://www.macsurrey.ac.uk/SystemQ/summary/">http://www.macsurrey.ac.uk/SystemQ/summary/</a></td>
</tr>
<tr>
<td>Surfboard</td>
<td>probably English only</td>
<td>pasted text</td>
<td><a href="http://www.giz.com/zinaria/surfboard/surfboard.dmg.gz">http://www.giz.com/zinaria/surfboard/surfboard.dmg.gz</a></td>
</tr>
<tr>
<td>SweSum</td>
<td>Danish, English, French, German, Spanish, Swedish</td>
<td>SDS (Web pages or pasted text)</td>
<td><a href="http://www.mada.kth.se/~martin/sweSum/index-eng.html">http://www.mada.kth.se/~martin/sweSum/index-eng.html</a></td>
</tr>
<tr>
<td>TextWise Content Repurposing Suite</td>
<td>probably English only</td>
<td>single document or e-mail</td>
<td><a href="http://www.textwise.com/technology/crs/demo.html">http://www.textwise.com/technology/crs/demo.html</a></td>
</tr>
</tbody>
</table>

Table 2.2: Some on-line demos of summarization systems, both commercial and academic
Characteristics of summary

- **Type**
  - Indicative vs informative
  - Extract vs Abstract (vs gist)
  - Generic vs query based
  - Background vs just-the-news
  - Single document vs multidocument
  - general vs domain dependent
  - textual vs multimedia

- **Input**
  - domain, genre, form, size
Characteristics of summary

• Related disciplines
  – IE, IR, Q&A, Topic identification (TI), Document Classification (DC), Document clustering, Event (topic) detection and tracking (TDT), Sentence simplification, Paraphrasing

• Evaluation

• Applications
  – biographies
  – Medical Reports summaries
  – E-mail summaries
    • Voice mails
  – Web pages summaries
  – News summaries
  – Headlines extraction
  – Support to IR
  – Meeting summaries
Characteristics of summary

• Current tasks involving research
  – Headline extraction
  – Cross-lingual MDS
  – Query-based
  – Definition Questions
  – Text simplification
  – Paraphrase generation
  – (query based) information needs
    • Summarization, definitional Q&A, ...
Dimensions for AS description

- Sparck Jones (1998)
  - Input
  - Purpose
  - Output
Outline

- Introduction
- **Problems of AS**
- Basic Architectures
- Evaluations
- Single-document Systems
- Multi-document Systems
- Headline extraction
- Query-based Systems
- Multimedia and Multilingual Systems
- Multitask Systems
• **Input**
  - Document structure
    - metadata
    - document organization
      - M.Y. Kan (2003), Elhadad et al, 2005
        - Tree based organization of medical reports
      - Teufel, Moens (2002)
        - organization of scientific articles
  - Multimedia integration
    - PERSIVAL, McKeown et al, 2001
Input 2

- **Input**
  - Domain
    - Domain dependent systems
    - Open domain systems
    - Adaptable systems
      - Columbia meta-summarizer
      - MULTIGEN: simple events
      - DEMS (bio configuration): biographies
        » Schiffman et al, 2001
      - DEMS (default configuration): otherwise
  - Detail level
    - ordinary, specialized, restricted
  - Restricted or free language (sublanguages)
Input

- Scale
  - Document length, granularity
  - news, articles, reports, books, ...
- Meaning units
  - segmentation
    - orthography or syntax
    - topic segmentation
  - type
    - conventional (paragraph, sentence, clause, ...)
    - specific (Semantic Content Unit, SCU, nugget, factoid)
- Media
  - text, video, meeting records, images, tables,
Input 4

- Input
  - Genres
    - Medical reports: Elhadad et al, 2005
    - Medical articles: Kan (2003)
    - Meeting records: Zechner (2001)
    - E-mails: Muresan et al (2001)
    - Voice mails (Koumpis, Renals, 2004)
  - Unit (SD vs MD)
  - Language ({mono, multi, cross}-lingual)
Purpose

- Situation
  - Complete Applications
  - Components of wider systems
    - MT, IR, QA
  - Task-driven vs General
- Audience
  - Background vs just-the-news
- Use
  - Retrieving, Kan et al (2001)
Output

- Content
  - generic vs user-focused vs query-driven

- Format
  - Plain text
  - list of items
  - structured

- Style
  - informative, indicative, aggregative, critic (Teufel, Moens, 2002)

- Production
  - extractive vs abstractive
  - use of NLG, Jing, Mc.Keown (2000)

- Relation with the source
  - Substitution, linked, included (highlighted text)

- Size
  - Compression degree
  - Headline
Approaches to AS

• Top-Down
  - query-driven
  - User needs
    • Some information
  - system
    • Specific needs for guiding the search
    • examples:
      - templates with slots with semantic features
      - list of relevant terms
  - Similar to IE

• Bottom-up
  - text-driven
  - User needs
    • Everything important
  - system
    • metrics over the generic importance (strategy)
    • Examples:
      - Connectivity degree in a semantic graph
      - frequency of (co-) occurrence of terms
  - Similar to IR
Basic schema

- Multi-document
- Single-document
- Query
- Conditions
- Summarizer
  - Extract
  - Abstract
  - Headline
Human Production

- Cremmins, 1982
  - Focus on basic characteristics of material to summarize
  - identify relevant information
  - Extract
  - organizing and reducing relevant information into coherent units, often one paragraph, and refine the summary by edition.

- Endres-Niggemeyer, 1989
  - Analitic rules (reduction and selection)
  - Synthetic rules (clarification, reorganization, style revision)
Tasks

- Extracting the most relevant fragments
  - sentences
  - paragraphs
  - passages
- ranking of these fragments by relevance trying to avoid redundancy
- Summary production
  - Saggion
    - text to text
    - propositions to text
    - information to text
Outline

• Introduction
• Problems of AS
• **Basic Architectures**
• Evaluations
• Single-document Systems
• Multi-document Systems
• Headline extraction
• Query-based Systems
• Multimedia and Multilingual Systems
• Multitask Systems
Techniques used

- 50s- 70s
  - Scientific documents
  - Statistical techniques
- 80s
  - AI techniques
- 90s
  - news
  - Hybrid systems
- 00s -
  - HE, MDS, qMDS, ...
Techniques used

• Processing level
  - Superficial
  - Entity
  - Discourse level

• Techniques used
  - Linguistic
  - Statistical
  - ML
  - IE
  - Text mapping and MT
  - Graph based
  - Combination
Techniques used

Superficial level

- Frequency of terms
- Position
  - Lead method
  - Title-based method
  - OPP (Optimum Position Policy)
    - Lin, Hovy, 1997
- Bias
  - Presence of terms from the title and headings
- cue words & phrases
Techniques used

Entity Level

- Build a representation of the document from their entities and relations
  - Words, MWT, word n-grams, NE
- Lexical Similarity
  - Distribucional or semantic
  - Budanitsky, Hirst, 2006
  - Rodríguez, 2003
- Proximity
- Logical relations
  - Agreement, contradiction, entailment
- Semantic relations
  - Predicate-argument
Techniques used

Entity Level

• Cohesion
  – Word Cooccurrence
  – Coreference
    • Baldwin, Morton, 1998
    • Bagga, Baldwin, 1998
  – Local saliency
  – Lexical chains
    • Barzilay, 1997
    • Fuentes, Rodriguez, 2002
  – MMR
    • Carbonell, Goldstein, 1998
Discoursive level

- Format and organization of the document
- Thread of topics
  - Topic Signatures
    - Hovy, Lin, 1999
    - Harabagiu, Lacatusu, 2005
  - Topic detection
- Topic Detection
  - Hovy, Lin, 1999,
  - Hovy, 2000
  - Radev et al, 2000, 2001
- Rethorical Structure
  - Marcu, 1997
  - Alonso, 2005
Linguistic techniques

- Sentence compression, reduction or simplification
  - (cut & paste)
    - Jing, 2000
    - Jing, McKeown, 1999
  - Knight, Marcu, 2000
  - Siddharthanan et al, 2004
Techniques used

Statistical techniques

• Language models
  – Berger, 2001,
  – Berger, Mittal, 2000

• Bayesian models
  – Kupiec et al, 1995
  – Schlesinger et al, 2001

• HMM

• ME, Logistic regression
  – Conroy et al, 2001
Techniques used

- Decision trees
  - Copeck et al, 2002
- ILP
  - Knight, Marcu, 2000
  - Tzoukerman et al, 2001
- SVM
  - Hirao et al, 2002
- Perceptron
  - Fisher, Roark, 2005, 2006
Techniques used

• Kan, McKeown, 1999
• Harabagiu, Lacatusu, 2002
• Daumé, 2002
• Lal, Rueger, 2002
Techniques used

Text mapping and MT

- Text mapping
  - Banko et al, 1999
  - Daumé III, Marcu, 2006
Techniques used

• Graphs of lexical connexion
  – (adjacency, grammatical relations, similarity, coreference)
  – Mani, Bloedorn, 1999
  – Leskovec et al, 2005

• Random Walk algorithms
  – LexRank - Erkan, Radev, 2004
  – TextRank - Mihalcea, Tarau, 2004

• Bigraphs
  – Zha, 2002
Techniques used

- Voting democratic or weighted
- ML
  - Classifiers
    - SVM, DT, NB,
- Examples
  - Kupiec et al, 1995
  - Hovy, Lin, 1999
  - Goldstein et al, 1999
  - Muresan et al, 2000
  - Kraaij et al, 2001
  - White et al, 2001
  - Teufel, Moens, 2002
  - Alonso, Fuentes, 2002
Outline

• Introduction
• Problems of AS
• Basic Architectures
• **Evaluations**
• Single-document Systems
• Multi-document Systems
• Headline extraction
• Query-based Systems
• Multimedia and Multilingual Systems
• Multitask Systems
Evaluation

• Process complex, controversial and costly

• **Basic Problem**
  – Very low agreement between human annotators

• Types of evaluation
  – intrinsic
    • Summary is evaluated per se or by comparing it to the source
      – coherence, cohesion
      – **informativity**
    – omissions
  – extrinsic
    • The system is evaluated against an (external) specific task
      – Can the summary be used instead of the document?
        » DC, Q&A
Evaluation

• Standard Metrics for evaluating extracts
  – Precision
  – Recall
  – Accuracy
  – F-score

• Specific Metrics
  – Based on ranking
    • RU (Relative Utility) (Radev)
  – Based on content
    • Overlap (n-gram, cosine, longest subsequence)
    • ROUGE (Lin, Hovy)
    • Basic Elements (Lin, Hovy)
    • Pyramid (Nenkova, Passonneau)
  – Based on combination
    • QARLA (Amigó)
extracted = a + b
relevant = a + d
**recall** = a / (a + d)
**precisión** = a / (a + b)
**accuracy** = (a + c) / (a + b + c + d)

\[
F = \frac{(\beta^2 + 1) \cdot p \cdot r}{\beta^2 \cdot p + r}
\]

\[
F_1 = \frac{2 \cdot p \cdot r}{p + r}
\]
Relative utility

- Radev, Tam, 2003
- N judges assign utility scores to each sentence of each cluster, First e are an extract of size e.

\[
\overrightarrow{U}_i = \{u_{i,1}, u_{i,2}, ..., u_{i,n}\} \\
= \text{sentence utility scores for judge } i \\
\text{for all } n \text{ sentences in the cluster}
\]

\[
\overrightarrow{U}_i' = \{\delta_{i,1} \cdot u_{i,1}, \delta_{i,2} \cdot u_{i,2}, ..., \delta_{i,n} \cdot u_{i,n}\} \\
= \text{extractive utility scores for judge } i
\]

- \(\delta_{i,j} = 1 \text{ for } j \leq e, = 0 \text{ for } j > e\)
Relative utility

\[ U_i = \sum_{j=1}^{n} u_{i,j} \]

= total self-utility for judge \( i \)

\[ U_i' = \sum_{j=1}^{n} \delta_{i,j} \cdot u_{i,j} \]

= total extractive self-utility for judge \( i \)

(computed over all \( n \) sentences)

\[ U_{i,k} = \sum_{j=1}^{n} \delta_{i,j} \cdot u_{k,j} \]

= total extractive cross-utility for \( i \) and \( k \) (\( i \neq k \))

\[ U_{i,avg} = \frac{1}{(N - 1)} \cdot \sum_{k=1}^{N} U_{i,k} \quad \text{for} \ i \neq k \]

= (non-symmetric) judge utility for judge \( i \).
Relative utility

\[ J \quad = \quad U_{avg} = \frac{1}{N} \cdot \sum_{i=1}^{N} U_{i,avg} \]

= interjudge performance

(average extractive cross-utility of all judges)

\[ U \quad = \quad \sum_{j=1}^{n} \sum_{i=1}^{N} u_{i,j} \]

= total extractive utility for all judges.

\[ U' \quad = \quad \sum_{j=1}^{n} \varepsilon_j \cdot \sum_{i=1}^{N} u_{i,j} \]

= total utility for all judges

\[ \sum_{j=1}^{n} \varepsilon_{i,j} = e. \]

\[ \varepsilon_j = 1 \text{ for } j \leq e, \quad = 0 \text{ for } j > e, \]
Relative utility 4

• For evaluating a summary of size \( e \) it is compared with those of the set of judges normalizing with the maxim score.

\[
S = \frac{\sum_{j=1}^{n} \delta_{s,j} \cdot \sum_{i=1}^{N} u_{i,j}}{U'}
\]

= system performance (\( \delta_{s,j} \) is equal to 1 for the top \( e \) sentences extracted by the system).

\[
R = \frac{1}{\binom{n}{e}} \sum_{t=1}^{\binom{n}{e}} S_t
\]

= random performance (computed over all \( \binom{n}{e} \) possible extracts of length \( e \)).

\[
D = \frac{S - R}{J - R} \text{ (normalized Relative Utility)}
\]
ROUGE

- Recall Oriented Understudy for Gisting Evaluation
- C.Y. Lin, E. Hovy, 2003
- Comparison with a gold standard (one or more judges)
- Overlapping units
  - ROUGE-n: n-gram co-occurrence statistics
  - ROUGE-L: Longest common substring.
  - ROUGE-W: Weighted longest common substring.
  - ROUGE-Sn: Skip-bigram co-occurrence statistics without gap length limit and with maximum gap lengths of n words.
  - ROUGE-SUn: Extension of skip-bigram.
Basic Elements

- Hovy, Lin, Zhou, 2005
- Takes into account head of phrases (nominal, verbal, adjectival and adverbial) and the relations between these heads and their arguments
- Matching at different levels
  - Word identity, lemma identity, synonymy
- Combination of scores
Pyramids

- Evaluates content and quality
- Pool of human summary as gold standard
- Summarization Content Units (SCU)
  - An SCU is a set of contributors expressing a unique content identified by a common label
- Starts from a set of manually built summaries
- A pyramid model that exploits the intersection and frequency of SCU is built.
- Informativity of a new summary against the pyramid model
- Matching of SCUs with the summary
  - Add weights of SCUs (Obs)
  - Compute Obs/Max
    - 2 scores original (precision) and modified (recall), depending on the value of Max
[Label: The Concorde crossed the Atlantic in less than 4 hours]

**Sum1**: making the transatlantic flight in 3 and ½ hrs

**Sum2**: The Concorde could make the flight in between New York and London or Paris in less than four hours

**Sum3**: completing its journey from London to New York in about 3 hours, 30 minutes

**Sum4**: took less than 4 hrs to cross the Atlantic
Building a Pyramid from Model Summaries (N=4)
Where $O_i$ is the number of SCUs of weight $i$, and $T_i$ represents each pyramid tier, $P = OBS/MAX$

$$OBS = \sum_{i=1}^{n} i \times O_i$$

$$MAX_M = \sum_{i=1}^{n} i \times |T_i|$$

where $j = \max_i$
On the European Community, for example, Mr Kinnock’s opposition to British membership was less that the EC was a ‘capitalist club’ and more that the hated Tories favoured it. When the referendum was held in 1975, just 18 per cent voted for a Welsh assembly. "Today's British government plays into the nationalist hand by a staggering lack of vision in adapting constitutionally as it has done over 150 years," he said.

The government plans to set up 21 new authorities to replace eight counties and 37 districts in Wales. A promise of a plebiscite on British participation in a single currency union would add extra assurance. The Welsh nationalists would probably be willing to provide ad hoc support to a Labour minority government, but without the sort of unconditional guarantees required for stable rule.

Yet with only days to the election to go, Mr Kinnock’s personality remains at the very centre of the British political plot. The prime minister also rejected a call for a Welsh parliament by Mr Ieuan Jones, the Plaid Cymru MP for Ynys Mon. Both Johns - Major and Smith - are against referenda. Pacts with Labour and the Tories are equally dangerous.

In 1992, the British prime minister, John Major, rejected a call for a Welsh parliament with devolved powers by a Welsh member of parliament. Major reminded him that results of an earlier referendum, in 1978, showed that only 18 percent of Welsh voters favored a parliament. In November 1993, the British Tory government pressed ahead with plans to reform the structure of local government in Wales by holding shadow elections to new unitary authorities as early as 1994.

It intended that these councils should take over in April 1995. This was in lieu of granting demands for an independent Welsh...
Amigó, 2006, Amigó et al, 2005, 206

Combination of metrics, human evaluators and systems

Three measures
  - QUEEN
    • Quality of a set of systems
  - KING
    • Quality of a set of metrics
  - JACK
    • Quality of a test set
Set $M$ of human models

Set $X$ of similarity metrics

Set $A$ of automatic systems
Set $M$ of human models

Set $X$ of similarity metrics

Set $A$ of automatic system

Queen$_{M,X}(a)$
Quality of System $a$

King$_{M,A}(X)$
Quality of metrics
• Document Understanding Conference
• Organized by NIST from 2001
• Tasks 2001 and 2002
  - SDS
    • fully automatic summarization of a single newswire/newspaper document in 100 words
  - MDS
    • given multiple newswire/newspaper documents on a single subject, to produce summaries of different lengths (400, 200, 100, 50 words in DUC 2001 and 200, 100, 50, 10 words in DUC 2002)
• Tasks 2003
  - Headline extraction
    • very short, 10-word, summaries from a single document.
  - Short (100-word) summaries focused by events
  - Given a cluster of documents produce short summaries focused by viewpoints
    • The viewpoint description was supposed to be a natural language string no larger than a sentence. It will describe the important facet(s) of the cluster the NIST assessor has decided to include in the short summary
  - Given a cluster of documents produce short summaries that answer a question
    • Given each document cluster, a question, and the set of sentences in each document deemed relevant to the question, the task consisted in the creation of short summaries of the cluster that answers the question.
DUC 3

• Tasks 2004
  - Task 1: very short ($\leq 75$ bytes) summary
  - Task 2: short ($\leq 665$ bytes) summary of cluster of documents
  - Task 3: very short cross-lingual summary
  - Task 4: short cross-lingual summary of document cluster
  - Task 5: short person profile
Tasks 2005 and 2006
- to synthesize from a set of 25-50 documents a brief, well-organized, fluent answer to a need for information that cannot be met by just stating a name, date, quantity, etc.
- This task models real-world complex question answering.
- Size: 250 words
<topic>
<num> d3011 </num>
<title> International Organized Crime </title>
<narr>
Identify and describe types of organized crime that crosses borders or involves more than one country. Name the countries involved. Also identify the perpetrators involved with each type of crime, including both individuals and organizations if possible.
</narr>
<granularity> specific </granularity>
</topic>

<topic>
<num> d426a </num>
<title> Law enforcement with dogs </title>
<narr>
What sorts of law-enforcement tasks are dogs being used for worldwide? What law enforcement agencies are using dogs? What breeds of dogs are being used?
</narr>
<granularity> general </granularity>
</topic>

Table 4.11: DUC 2005 complex natural language questions.
Outline

• Introduction
• Problems of AS
• Basic Architectures
• Evaluations

• **Single-document Systems**
• Multi-document Systems
• Headline extraction
• Query-based Systems
• Multimedia and Multilingual Systems
• Multitask Systems
Lexical Chains

• Regina Barzilay (1997)
• Also used for
  – hyperlinks generation
    • (Green, 1997, 1999)
  – Topic segmentation
    • (Marti A. Hearst, 1994)
  – Topic identification
    • (Lin, 1998)
  – AS
    • (Barzilay, 1997), (Brunn et al, 2001), (Fuentes, Rodríguez, 2002), (Stokes et al, 2004)
• Factors for defining the chain
  – Word cooccurrence, coreference, relevance of words, semantic relations, …
• Fuentes, Rodríguez, 2002 use a relative classification
• \( \mu_S \) is the mean of scores of all the chains
• \( \sigma_S \) standard deviation

\[
\tau = \mu_S + 2 \cdot \sigma_S
\]

\[
\text{Strong} = \{ c \mid \text{score}_c \geq \tau \}
\]

\[
\text{Medium} = \{ c \mid \tau > \text{score}_c \geq \tau / 2 \}
\]

\[
\text{Light} = \{ c \mid \tau / 2 > \text{score}_c \}
\]
Lexical Chains

Figure 4.1: LCsum summarization process.
**Lexical Chains**

<table>
<thead>
<tr>
<th>unit₁</th>
<th>unit₂</th>
<th>...</th>
<th>unit$_{N-1}$</th>
<th>unit$_N$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- chain₁
- chain₂
- chain₃
- chain$_{M-1}$
- chain$_M$

**weight of the unit**

**weight of the chain**
Lexical Chains

3 types of chains:
- Lexical chains
- Coreference chains
- NE chains
RST states that the elements in the text are related by rhetorical relations supporting it. The basic idea is that the text can be organized into text spans, expressing the goals of talkers related by these rhetorical relations. Mann, Thomson propose 25 rhetorical relations.
• Construction of RTS tree
• RST parsing (usually tied to recognition of Rethorical markers)
• Application of measures of cohesion (e.g. vocabulary overlap) for connecting the trees
• Alonso, 2005
  • \textit{dm} neutral, \textit{nuc} relevante, \textit{sat} irrelevante
  • \textit{r} right directed, \textit{b} both sides
[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 (Marcu)

- 2 Elaboration
  - 2 Background Justification
  - 3 Elaboration
    - 4 Contrast
      - 4 Evidence Cause
  - 5 Elaboration
    - 7 Concession
    - 8 Example
      - 9 Antithesis
    - 10 Example

Summarization = selection of the most important units

2 > 8 > \{3, 10\} > \{1, 4, 5, 7, 9\} > 6
Statistical methods

• Bayesian models
  – Schlesinger, Baker, 2001
  – Daumé III, Marcu, 2005

• HMM
  – Conroy, Schlesinger, ... 2001, ... , 2006

• Logistic Regression
  – Conroy, Schlesinger, ... 2001,

• Language Models

• Noisy Channel Model
  – Berger 2001, Berger Mittal, 2000
Statistical methods

Conroy et al, 2001

- Naive Bayes
- HMM
- Logistic regression
- features:
  - posición en la oración, # tokens, # query terms en la oración, distancia a los query terms

Kraaij et al, 2001, 2002

- Combination of a mixture LM
  - content-based (smoothed unigram)
- and Naive Bayes
  - non content features (sentence length, 1st sentence, ...) combined using odds of salience
Noisy Channel Model Approach
- OCELOT
  - summarizing Web pages
  - Learning from Open Directory Project
  - Content selection (models of gisting), word ordering, searching
- Learning from FAQ files
  - query-relevant summarization
  - relevance vs fidelity
  - shrinkage for smoothing

Berger, 2001, Berger, Mittal, 2000
• Conroy, Schlesinger, O’leary, 2001, …, 2006
  – IDA, Univ. Maryland
• Clustering Linguistic and Statistics for Summarization Yield
• MDS
  – Sentence splitting & trimming
  – Sentence scoring
  – Sentence selection for redundancy removal and improving flow
• 50% terms in human summaries do not occur in document(s)
• Agreement in human summarizers vocabulary 40%
• Human variation: unigram bow
• Topic $\tau$
• Probability a human includes $t$ in a summary $P(t|\tau)$
• For a sentence $x$

$$\omega(x) = \frac{1}{|x|} \sum_{t \in T} x(t)P(t|\tau)$$

• Where $x(t) = 1$ if $t$ occurs in $x$ and zero otherwise
Learning $P(t|\tau)$

- MLE from $h$ sample summaries
  - $C_{it}(\tau) = 1$ if $i$-th summary contains $t$
- Oracle score $\tilde{\omega}(x)$ computed from the estimate of $P$
- Another approximation
  - $\tau$ topic description $\rightarrow$ query terms
  - collection of relevant documents $\rightarrow$ signature terms
- Improvements
  - Including NEs as third set of terms
    - NO IMPROVEMENT
  - Pseudo-relevance feedback
• Pseudo-relevance feedback
  - 1) compute $\omega_{qs}$ for each sentence
  - 2) select $k$ tops

• A term-sentence matrix
  - $k$ tops correspond to columns $j_1, \ldots, j_k$

• $\rho(t)$ expectation that a term is used in the extract of these $k$ sentences

\[ \rho(t) = \frac{1}{k} \sum_{i=1}^{k} A(t, j_i). \]

\[ P_{qsp}(t|\tau) = \frac{1}{2} P(t|\tau) + \frac{1}{2} \rho(t). \]

\[ \omega_{qsp}(x) = \frac{1}{|x|} \sum_{t \in T} x(t) P_{qsp}(t|\tau) \]
• Reducing the redundancy of the selected sentences
• Most popular method: MMR
• They use **pivoted QR** until DUC-2005, in DUC-2006 non-negative-QR decomposition
  - Conroy et al, 2001
• Gram-Schmidt orthogonalization method
• **B** term-sentence matrix
  - T columns, m rows
  - \( B_{ij} = 1 \) if j contains i
• \( A_w \) normalized
  - Each column is normalized dividing each element by its 2-norm (euclidean distance)
• for \( i = 1, \ldots, \min(m,T) \)
  - choose column l that has the **highest norm** (\( a_l \))
  - set \( \text{Index}_i = l \)
  - set \( q_i = a_l / \| a_l \| \)
  - **update the other columns to make them orthogonal** to \( a_l \)
    - for each unchosen column \( a_j \)
      - \( r_{ij} = a_j^T q_i \)
      - \( a_j = a_j - r_{ij} q_i \)
• The summary of length k will contain \( \text{Index}_i, \ldots, \text{Index}_k \)
The idea of pivoted QR is in every iteration, after the selection is made, subtracting off from all the remaining columns the components that are in the direction of the selected column.
Graph-based measures of importance

- Mihalcea, Radev tutorial on RANLP 2005
- Centrality, Relevance, Saliency of nodes
  - paragraphs, sentences, words, ...
- Several forms of representation
  - graphs (undirected)
  - digraphs
  - weighted or binary
  - bigraphs
Graph-based measures of importance

• Random Walk Algorithms
  - Decide the importance of a vertex within a graph
  - A link between two vertices = a vote
    - Iterative voting => Ranking over all vertices
  - Model a random walk on the graph
    - A walker takes random steps
    - Converges to a stationary distribution of probabilities
    - Guaranteed to converge if the graph is
      - Aperiodic – holds for non-bipartite graphs
      - Irreducible – holds for strongly connected graphs
  - Examples
    • PageRank (Brin, Page, 1998)
    • HITS (Kleinberg, 1999)
Graph-based measures of importance

- **PageRank**
  - Random walks over the Web following hyperlinks
  - some probability of jumping randomly out of the page
  - The fraction of time spent in each page is the PageRank

- **HITS (Hyperlinked Induced Topic Search)**
  - Defines 2 types of importance:
    - **Authority**
      - Can be thought of as being an expert on the subject
      - “You are important if important hubs point to you”
    - **Hub**
      - Can be thought of as being knowledgeable about the subject
      - “You are important if important authorities point to you”
  - Hubs and Authorities mutually reinforce each other
  - random walk where:
    - Odd steps you walk from hubs to authorities (follow outlink)
    - Even steps you walk from authorities to hubs (follow inlink)
Graph-based measures of importance

- **PageRank**
  - simple
    \[
    PR(V_i) = (1 - d) + d \sum_{j \in \text{In}(V_i)} \frac{1}{|\text{Out}(V_j)|} PR(V_j)
    \]
    \[d \in [0,1]\]
  - weighted graphs
    \[
    PR^w(V_i) = (1 - d) + d \sum_{j \in \text{In}(V_i)} \frac{w_{ji}}{\sum_{V_k \in \text{Out}(V_j)} \sum_{V_k \in \text{Out}(V_j)} w_{jk}} PR^w(V_j)
    \]

- **HITS**
  \[
  HITS_H(V_i) = \sum_{j \in \text{Out}(V_i)} HITS_A(V_j)
  \]
  \[
  HITS_A(V_i) = \sum_{j \in \text{In}(V_i)} HITS_H(V_j)
  \]
Salton et al, 1999

- Salton, Singhal, Mitra, Buckley, 1999
- Representing documents and paragraphs using VSM
- Similarity: escalar product
- Representing document as undirected graphs
  - Nodes = paragraphs
  - Edges weighted with sim
- Salient nodes are those connected with many others with sim over a threshold

\[ D_i = (d_{i1}, \ldots, d_{in}) \]

\[ \text{sim}(D_i, D_j) = \sum d_{ik} \cdot d_{jk} \]
TextRank

- Mihalcea & Tarau, 2004
- Graph-based ranking algorithms for extractive summarization
- Weighted Graph representation
  - nodes: sentences
  - edges: similarity between sentences
- Run a Graph-based ranking algorithm
  - $\text{PR}^w$, $\text{HITS}_A^w$, $\text{HITS}_H^w$
- Get the N best ranked sentences
- MDS
  - cascaded summarization
  - using SDS for each document
  - using SDS for performing summary of summaries
LexRank

- Erkan, Radev, 2004
- Idea: Sentences more similar to others are more central (salient)
- Centroid-based
  - 3 measures
    - Degree
    - LexRank with threshold
    - Continuous LexRank
- Sentence similarity

\[
\text{idf-modified-cosine}(x, y) = \frac{\sum_{w \in x,y} \text{tf}_{w,x} \text{tf}_{w,y} (\text{idf}_w)^2}{\sqrt{\sum_{x_i \in x} (\text{tf}_{x_i,x} \text{idf}_{x_i})^2} \times \sqrt{\sum_{y_i \in y} (\text{tf}_{y_i,y} \text{idf}_{y_i})^2}}
\]
LexRank

- Similarity graph (SG)
  - Nodes: sentences
  - Edges: similarity scores
- Degree centrality = degree of a sentence in SG
- SG is represented by an adjacency matrix $A$
- $A$ can be normalized as $B$

\[
B(i, j) = \frac{A(i, j)}{\sum_k A(i, k)}
\]
Iraqi Vice President Taha Yassin Ramadan announced today, Sunday, that Iraq refuses to back down from its decision to stop cooperating with disarmament inspectors before its demands are met.

Iraqi Vice president Taha Yassin Ramadan announced today, Thursday, that Iraq rejects cooperating with the United Nations except on the issue of lifting the blockade imposed upon it since the year 1990.

Ramadan told reporters in Baghdad that "Iraq cannot deal positively with whoever represents the Security Council unless there was a clear stance on the issue of lifting the blockade off of it.

Baghdad had decided late last October to completely cease cooperating with the inspectors of the United Nations Special Commission (UNSCOM), in charge of disarming Iraq's weapons, and whose work became very limited since the fifth of August, and announced it will not resume its cooperation with the Commission even if it were subjected to a military operation.

The Russian Foreign Minister, Igor Ivanov, warned today, Wednesday against using force against Iraq, which will destroy, according to him, seven years of difficult diplomatic work and will complicate the regional situation in the area.

Ivanov contended that carrying out air strikes against Iraq, who refuses to cooperate with the United Nations inspectors, “will end the tremendous work achieved by the international group during the past seven years and will complicate the situation in the region.”

Nevertheless, Ivanov stressed that Baghdad must resume working with the Special Commission in charge of disarming the Iraqi weapons of mass destruction (UNSCOM).

The Special Representative of the United Nations Secretary-General in Baghdad, Prakash Shah, announced today, Wednesday, after meeting with the Iraqi Deputy Prime Minister Tariq Aziz, that Iraq refuses to back down from its decision to cut off cooperation with the disarmament inspectors.

British Prime Minister Tony Blair said today, Sunday, that the crisis between the international community and Iraq “did not end” and that Britain is still “ready, prepared, and able to strike Iraq.”

In a gathering with the press held at the Prime Minister's office, Blair contended that the crisis with Iraq “will not end until Iraq has absolutely and unconditionally respected its commitments” towards the United Nations.

A spokesman for Tony Blair had indicated that the British Prime Minister gave permission to British Air Force Tornado planes stationed in Kuwait to join the aerial bombardment against Iraq.
LexRank 4

Edge Weights:
- [0.3,1.0] (black)
- [0.2,0.3] (dark gray)
- [0.1,0.2] (light gray)
- [0.0,0.1] (lightest gray)
LexRank 5
LexRank

- Distribute centrality between neighbours
  - $p = B^T p$
  - $p^T B = p^T$

- $p^T$ is the left eigenvector of $B$ corresponding to the eigenvalue 1
- $B$ has to be irreducible and aperiodic
LexRank

- Following PageRank (Page et al, 1998)
  - Lexical PadeRank
  - Reserving some low probability of jumping to any node in SG

\[
p(u) = \frac{d}{N} + (1 - d) \sum_{v \in \text{adj}[u]} \frac{p(v)}{\text{deg}(v)}
\]

\[
p = [dU + (1 - d)B]^T p
\]
LexRank

- Using a weighted graph
- Continuous LexRank

\[
p(u) = \frac{d}{N} + (1 - d) \sum_{v \in \text{adj}[u]} \frac{\text{idf-modified-cosine}(u, v)}{\sum_{z \in \text{adj}[v]} \text{idf-modified-cosine}(z, v)} p(v)
\]

LexNet
Radev et al, 2006
LexRank

- Topic sensitive LexRank
- Ottebacher et al, 2006
- Relevance to the query
- Computing $\text{idf}_w$ for each word (after stemming) in the set

\[ \text{idf}_w = \log \left( \frac{N + 1}{0.5 + sf_w} \right) \]

- Relevance of a sentence

\[ \text{rel}(s|q) = \sum_{w \in q} \log(tf_{w,s} + 1) \times \log(tf_{w,q} + 1) \times \text{idf}_w \]
LexRank

- Mixture model
  - Similarity

\[
\text{sim}(x, y) = \frac{\sum_{w \in x, y} \text{tf}_{w,x} \text{tf}_{w,y} (\text{idf}_w)^2}{\sqrt{\sum_{x_i \in x} (\text{tf}_{x_i,x} \text{idf}_{x_i})^2} \times \sqrt{\sum_{y_i \in y} (\text{tf}_{y_i,y} \text{idf}_{y_i})^2}}
\]

- Relevance of s

\[
p(s|q) = d \frac{\text{rel}(s|q)}{\sum_{z \in C} \text{rel}(z|q)} + (1-d) \sum_{v \in C} \sum_{z \in C} \frac{\text{sim}(s, v)}{\text{sim}(z, v)} p(v|q)
\]

\[
p = [dA + (1 - d)B]^T p
\]

- When A is uniform and B normalized binary the algorithm reduces to LexRank
Zha, 2002

- Generic Text Summarization + Keyphrase extraction
- SDS or MDS
- Mutual Reinforcement Principle
- Unsupervised approach
- Approach
  - 1) Clustering sentences into topical groups
  - 2) Within each topical group select keyphrases and sentences by saliency score
• **Bigraph representation of a document**
  - Set of terms $T = \{t_1, \ldots, t_n\}$
  - Set of sentences $S = \{s_1, \ldots, s_m\}$
  - Weighted bipartite graph
    - Nodes: $T \cup S$
    - Edges
      - $(t_i, s_j)$ is weighted by the number of times $t_i$ occur in $s_j$
    - $G(T,S,W)$

• **Saliency**
  - A term should have a high salience if it appears in many sentences with high salience while a sentence has a high salience if it contains many terms with high salience
• Salience of a term

\[ u(t_i) \propto \sum_{v(s_j) \sim u(t_i)} w_{ij} \cdot v(s_j) \]

• Salience of a sentence

\[ v(s_j) \propto \sum_{v(s_j) \sim u(t_i)} w_{ij} \cdot u(t_i) \]

• Salience of terms

\[ u = \frac{1}{\sigma} Wv \]

• Salience if sentences

\[ v = \frac{1}{\sigma} W^Tu \]
• For $\sigma = \text{Max eigenvalue of } W$ components of $u$ and $v$ (eigen vectors) are non negative
• Max $\{u, \sigma, v\}$ over $W$
• Iteration till convergence of equations $u$ and $v$
• On convergence $\sigma$ is computed
Outline

- Introduction
- Problems of AS
- Basic Architectures
- Evaluations
- Single-document Systems
- Multi-document Systems
- Headline extraction
- Query-based Systems
- Multimedia and Multilingual Systems
- Multitask Systems
Multidocument Summarization (MDS)

Goals

• Content of a collection of documents
• Briefing
  – concise summary of the factual matter of a set of news articles on the same or related events (SUMMONS, Radev, 1999)
• Information updating
• Looking for user information needs in a collection of documents
MDS 2

Differences SDS MDS

- Lower compression level
- anti-redundancy
- temporal dimension
- Coreference more challenging
Requirements

- Clustering of documents and passages
  - Similarity measures between units (documents, passages, paragraphs, sentences)
- recall
- anti-redundancy
- Summary cohesion
- quality
  - readable
  - relevant
  - context
- Sources inconsistencies
- updating
Types

• Common sections of documents of the collection
• Common sections + unique sections
• Centroids
• Centroids + outliers
• Last document + outliers
• Common sections + unique sections + temporal weighting score
• **Information Fusion** (Barzilay et al, 1999)
  - articles presenting different descriptions of the same news
  - Repetition is a good indicator of relevance
  - automatically generate a concise summary by identifying similarities and differences across a set of related documents.
    - Identifying **themes**
    - **Information Fusion**
    - **Generation** (**Reformulation**)
Techniques

- Most are the same used in SDS
- Clustering
- IE
  - Context modeling
    - (Lal, Rueger, 2002), (Kan, McKeown, 1999)
  - Templates adapted to domains or events
    - (Harabagiu, Lacatusu, 2002), (Daumé et al, 2002)
- Centroid-based
  - (Radev et al, 2004), (Saggion, Gaizauskas, 2004)
- Graph-based
  - (Radev et al, 2004), (Erkan, Radev, 2004), (Mihalcea, Tarau, 2004), (Zha, 2002)
- SVM
  - (Hirao et al, 2003)
Outline

- Introduction
- Problems of AS
- Basic Architectures
- Evaluations
- Single-document Systems
- Multi-document Systems
- **Headline extraction**
- Query-based Systems
- Multimedia and Multilingual Systems
- Multitask Systems
Headline Extraction

- A headline is a highly concise representation of the most relevant points contained in a document. It can consist of a sentence, either extracted from the document or automatically generated, or, sometimes, of a list of relevant terms.
- The main characteristic of a headline is its extremely small length (usually between 5 and 10 words).
- 3 main tasks
  - Identification of relevance signaling words.
  - Combination of the results proposed by simpler individual methods.
  - Some form of compression or simplification of the extracted sentences.
Headline Extraction

- **Identification of relevance signaling words**
  - (Schlessinger et al 2001)
    - lead property, specificity of verbs, use of Concept Sets derived from WN
  - (Kraaij et al, 2002)
    - identify the most informative topical NP:
      - Sentences are ranked by an hybrid model that merges i) a unigram LM, mixing cluster and document models, for scoring sentences and ii) a Bayesian model on content features like cue phrases, position, length, etc. A Trigger Word Pool is built from the ranked sentences and the most salient trigger word is selected.
      - Then, all maximal NPs containing the trigger word are found, and the one in the highest ranked sentence is taken as a headline.
  - (Daumé et al, 2002)
    - go beyond words or NP and perform a full parsing of the document for getting the main entities and their relations. From them the headline is built.
  - (Zajik et al, 2002)
    - Use a Noisy Channel Model (NCM), implemented by means of a HMM, for selecting a stream of headline words from a stream of story words. The first is modelled by a bigram LM and the latter by unigram one. The headline is generated by a simple Viterbi decoder constrained by the model. A set of penalty constraints (length, position, gap and string) is added.
Headline Extraction

- Combination of the results proposed by simpler individual methods
  - (Lin, 1999)
    - linear combination
  - (Aone et al, 1997)
    - DT
  - MEAD (Erkan, Radev, 2004)
    - combining centroid, position and length

- Some form of compression or simplification of the extracted sentences
  - (Knight, Marcu, 2000)
    - translation from full text to summary using NCM or DT
  - (Lal, Rueger, 2002)
    - lexical simplification
  - TOPIARY (Zajik et al, 2004), (Dorr et al, 2003)
    - sentence compression
Outline

- Introduction
- Problems of AS
- Basic Architectures
- Evaluations
- Single-document Systems
- Multi-document Systems
- Headline extraction
- **Query-based Systems**
- Multimedia and Multilingual Systems
- Multitask Systems
Query-based Summarization

- User information need expressed with a query
  - NL question
  - IR query
  - DUC 2005 and 2006

- Two tasks
  - Identification of relevant sentences
  - Similarity between sentence and query
Query_based Summarization

- Identification of relevant sentences
  - frequency
    - (Nenkova, Vanderwende, 2005)
  - oracle score
    - (Conroy et al, 2005)
  - Bayes
    - (Daumé, Marcu, 2005), (Daumé, 2006)
  - Perceptron (sentence ranking)
    - (Fisher, Roark, 2006)
  - Centroid
  - Graphs

- Similarity between sentence and query
Query_based Summarization

- **Similarity between sentence and query**
  - Tree similarity
    - (Schilder, McInnes, 2006)
  - Q&A
    - (Fuentes et al, 2005), (Lacatusu et al, 2006), (McKeown et al, 2006), (Mollá, 2006)
  - Enriched VSM
    - Query expansion
      - (Alfonseca et al, 2006)
    - Syntactic and semantic relations
      - (Lacatusu et al, 2006)
    - LSA
      - (Miller, 2003), (Hachey et al, 2005)
Query_based Summarization

- Combina varias estrategias para
  - descomposición de la pregunta
  - generación del resumen
- Question Processing Module
  - descomposición de preguntas complejas en simples
    - extracción de keywords
    - descomposición sintáctica
    - descomposición semántica
- Uso de un sistema convencional de Q&A
  - Harabagiu et al, 2005
- Sistema de MDS para obtener los 6 mejores resúmenes
- Sistema de textual entailment para elegir el mejor respecto a la query
  - uso del pyramid modified score para evaluar cada candidato

GISTexter, Lacatusu et al, 2006
Query-based Summarization

QASUM-TALP

M. Fuentes (2008)
Query_based Summarization

What law enforcement agencies are using dogs?

Generated Questions

When agencies are using dogs?
Where agencies are using dogs?
Who agencies are using dogs?

Error due to misclassification of agencies as human

M. Fuentes (2008)
Outline

• Introduction
• Problems of AS
• Basic Architectures
• Evaluations
• Single-document Systems
• Multi-document Systems
• Headline extraction
• Query-based Systems
• **Multimedia and Multilingual Systems**
• Multitask Systems
Outline

• Introduction
• Problems of AS
• Basic Architectures
• Evaluations
• Single-document Systems
• Multi-document Systems
• Headline extraction
• Query-based Systems
• Multimedia and Multilingual Systems
• **Multitask Systems**