
- Print, film, magnetic, and optical storage media produced about 5 exabytes ($10^{18}$) of new information in 2002. 92% of the new information was stored on magnetic media, mostly in hard disks.
  - If digitized with full formatting, the 17 million books in the Library of Congress contain about 135 terabytes of information;
  - 5 exabytes of information is equivalent in size to the information contained in 37,000 new libraries the size of the Library of Congress book collections.
- The amount of new information stored on paper, film, magnetic, and optical media has about doubled during 1999 – 2002.
- Information flows through electronic channels – telephone, radio, TV, and the Internet – contained almost 18 exabytes of new information in 2002, three and a half times more than is recorded in storage media (98% are telephone calls).

* Source: How much information 2003

The Big & Bigger Search Engines

<table>
<thead>
<tr>
<th>Search Engine</th>
<th>Reported Size</th>
<th>Page Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>8.1 billion</td>
<td>101K</td>
</tr>
<tr>
<td>MSN</td>
<td>5.0 billion</td>
<td>150K</td>
</tr>
<tr>
<td>Yahoo</td>
<td>4.2 billion</td>
<td>500K</td>
</tr>
<tr>
<td>Ask Jeeves</td>
<td>2.5 billion</td>
<td>101K</td>
</tr>
</tbody>
</table>

* Nov, 2004 searchenginewatch.com

Billions Of Textual Documents Indexed (6, 2002 - 9, 2003)
Got Spams?

- On Spam: Wasting time on the Internet (3/25/98) by Bill Gates
  - "Wasting somebody else's time strikes me as the height of rudeness. We have only so many hours, and none to waste."
- Why I Hate Spam (6/23/2003) by Bill Gates
  - "SEATTLE -- Like almost everyone who uses e-mail, I receive a ton of spam every day. Much of it offers to help me get out of debt or get rich quick. It would be funny if it weren't so irritating."
  - "Already, spam filters built into MSN and Hotmail servers block 2.4 billion messages a day before they reach subscribers' inboxes."

Which School is Better?

Is this Shoe Good or Bad?
Which Movie for Friday Night?

Plot Summary for
Star Wars: Episode III - Revenge of the Sith (2005)

June 2, 2005

Browser a Summarized Web

Solution: Summarization

- Informed decision through summarization
- What is summarization?
  - A summary is a concise restatement of the topic and main ideas of its source.
    - Concise – giving a lot of information clearly and in a few words. (Oxford American Dictionary)
    - Restatement – in your own words.
    - Topic – what is the source about?
    - Main ideas – important facts or arguments about the topic.
Kinds of Summaries

- Based on two different perspectives
  - The author's point of view
    - Paraphrase summary (indicative vs. informative)
      - In the fall of 1978, Fran Tate wrote some hot checks to finance the opening of a Mexican restaurant in Barrow, Alaska. All the banks she had applied to for financing had turned her down. But Fran was sure that Barrow lusted for a Mexican restaurant, so she went ahead with her plans. (“In Alaska: Where the Chili is Chilly”, by Gregory Jaynes)
  - The summary writer’s point of view
    - Analytical summary
      - For Gregory Jaynes, Fran Tate embodies the astonishing independence and toughness that is typical of Alaska at its best. Tate has opened a Mexican restaurant—an unlikely enterprise—in Barrow, Alaska.

How to Summarize?

- Topic Identification
  - Find the most important information—the topic.
    - The Egyptian civilization was one of the most important cultures of the ancient world. This civilization flourished along the rich banks and delta of the Nile River for many centuries, from 3200 B.C. until the Roman conquest in 30 B.C. …*
  - Find main ideas that support the topic and show how they are related.
    - Among the many contributions made by the Egyptian culture is the hieroglyphic writing system they invented. This is one of the earliest writing systems known and it was used from about 3200 B.C. until 350 A.D. Hieroglyphics are a form of picture writing, in which each symbol stands for either a single object or a single sound. The symbols can be combined into long strings to form words and sentences. Other cultures, such as the Hittites, the Cretans, and the Mayans, also developed picture writing, but these systems are not related to the Egyptian system, nor to one another.*

- Topic Interpretation (I)
  - Combine several main ideas into a single sentence.
    - Among the many contributions made by the Egyptian culture is the hieroglyphic writing system they invented. This is one of the earliest writing systems known and it was used from about 3200 B.C. until 350 A.D. Hieroglyphics are a form of picture writing, in which each symbol stands for either a single object or a single sound. The symbols can be combined into long strings to form words and sentences. Other cultures, such as the Hittites, the Cretans, and the Mayans, also developed picture writing, but these systems are not related to the Egyptian system, nor to one another.*

- Topic Interpretation (II)
  - Substitute a general form for lists of items or events.
    - List of items: John bought some milk, bread, fruit, cheese, potato chips, butter, hamburger meat
      - John bought some groceries.*
    - List of events: A lot of children came and brought presents. They played games and blew bubbles at each other. A magician came and showed them some magic. Later Jennifer opened her presents and blew out the candles on her cake.*
      - Jennifer had a birthday party.*
  - Remove trivial and redundant information
    - We have gotten word of an enormous deal in the publishing deal. Hillary Clinton, senator-elect has accepted an $8 million advance to publish her book. That is, according to the New York Times, the largest advance ever for a first-time author. Hillary Clinton has accepted an advance of $8 million for her book. Senator-elect Hillary Clinton’s advance is reported to be $8 million for her book. Senator-elect Hillary Clinton Friday agreed to sell Simon & Schuster a memoir of her years as first lady, for the near-record advance of about $8 million. Hillary Rodham Clinton Friday night agreed to sell Simon & Schuster a memoir of her years as first lady, for the near-record advance of around $8 million. Senator-elect Hillary Rodham Clinton Friday night agreed to sell Simon & Schuster a memoir of her years as first lady, for the near-record advance of about $8 million. Senator-elect Hillary Rodham Clinton Friday night agreed to sell Simon & Schuster a memoir of her years as first lady, for the near-record advance of about $8 million. Senator-elect Hillary Rodham Clinton Friday night agreed to sell Simon & Schuster a memoir of her years as first lady, for the near-record advance of about $8 million. Senator-elect Hillary Rodham Clinton Friday night agreed to sell Simon & Schuster a memoir of her years as first lady, for the near-record advance of about $8 million.
      - Senator-elect Hillary Rodham Clinton Friday night agreed to sell Simon & Schuster a memoir of her years as first lady, for the near-record advance of about $8 million.
  - Examples adapted from Summary Street®: http://lsa.colorado.edu/summarystreet
Aspects of Summarization

• Input
  (Sparck Jones 97, Hovy and Lin 99)
  - Single-document vs. multi-document...fuse together texts?
  - Domain-specific vs. general...use domain-specific techniques?
  - Genre...use genre-specific (newspaper, report...) techniques?
  - Scale and form...input large or small? Structured or free-form?
  - Monolingual vs. multilingual...need to cross language barrier?

• Purpose
  - Situation...embedded in larger system (MT, IR) or not?
  - Generic vs. query-oriented...author’s view or user’s interest?
  - Indicative vs. informative...categorization or understanding?
  - Background vs. just-the-news...does user have prior knowledge?

• Output
  - Extract vs. abstract...use text fragments or re-phrase content?
  - Domain-specific vs. general...use domain-specific format?
  - Style...make informative, indicative, aggregative, analytic...

Table of contents

1. Motivation and Introduction.
2. Approaches and paradigms.
3. IE and Summarization methods.
4. Evaluating summaries.
5. The future.

Two Psycholinguistic Studies

• Coarse-grained summarization protocols from professional summarizers (Kintsch and van Dijk, 78):
  - Delete material that is trivial or redundant.
  - Use superordinate concepts and actions.
  - Select or invent topic sentence.

• 552 finely-grained summarization strategies from professional summarizers (Endres-Niggemeyer, 98):
  - Self control: make yourself feel comfortable.
  - Processing: produce a unit as soon as you have enough data.
  - Info organization: use “Discussion” section to check results.
  - Content selection: the table of contents is relevant.
Computational Approach: Basics

**Top-Down:**
- I know what I want! — don't confuse me with nonsense!
- User wants 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 wants anything that's important.
- System needs generic importance metrics, used to rate content.

Top-Down: Info. Extraction (IE)

- **IE task:** Given a form and a text, find all the information relevant to each slot of the form and fill it in.
- **Summ-IE task:** Given a query, select the best form, fill it in, and generate the contents.
- **Questions:**
  1. IE works only for very particular forms; can it scale up?
  2. What about info that doesn't fit into any form—is this a generic limitation of IE?

Bottom-Up: Info. Retrieval (IR)

- **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?
Paradigms: IE vs. IR

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

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

Deep and Shallow, Down and Up...

**Today:**
Increasingly, techniques hybridize: people use word-level counting techniques to fill IE forms' slots, and try to use IE-like discourse and quasi-semantic notions in the IR approach.

**Thus:**
You can use either deep or shallow paradigms for either top-down or bottom-up approaches!

Toward the Final Answer...

**Problem**: What if neither IR-like nor IE-like methods work?
- sometimes counting and forms 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).

**Word counting**

Mrs. Coolidge: "What did the preacher preach about?"
Coolidge: "Sin."
Mrs. Coolidge: "What did he say?"
Coolidge: "He's against it."

**Inference**
The Optimal Solution...

Combine strengths of both paradigms...

...use IE/NLP when you have suitable form(s),
...use IR when you don’t...

...but how exactly to do it?

Table of contents

1. Motivation and Introduction.
2. Approaches and paradigms.
3. IE and Summarization methods.
4. Evaluating summaries.
5. The future.

Information Extraction Method

- **Idea**: content selection using forms (templates)
  - Predefine a form, 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 form.
  - Generate the content of the form as the summary.

- **Previous IE work**:
  - FRUMP (DeJong, 79): ‘sketchy scripts’ of terrorism, natural disasters, political visits...
  - McFRUMP (Mauldin, 89): forms for conceptual IR.
  - SCISOR (Rau and Jacobs, 90): forms for business.
  - SUMMONS (McKeown and Radev, 95): forms for news.
  - GaTexter (Harabagiu et al., 03): dynamic templates based on WordNet.
IE-Based Summarization

- **Topic Identification**
  - A template defines the topic of interest and important facts associated with the topic. This also defines the coverage of IE-based summarizer.
  - Problem: how to select a relevant template? (i.e. the frame initiation problem)

- **Topic Interpretation**
  - Problem: how to fill template slots and merge multiple templates?

- **Generation**
  - Problem: how to generate summaries from the filled and merged templates?

FRUMP (DeJong 1979)

- **Domain:** General news
- **Input:** UPI news stories
- **Output:** summaries (single doc)
- **Template:** 60 sketchy scripts that contain important events in a situation.
  - Meeting, earthquake, theft, war, ...
  - Represented as conceptual dependency (CD)
    - Encode constraints on script variables, between script variables, and between scripts.
    - Ex: John went to New York,
      

```
(\text{ACTOR} (*John*)
 \iff (\text{PTRANS} *)
 \text{OBJECT} (*JOHN*)
 \text{TO} (*NEW YORK*)
```

- **Evaluation**
  - 6 days run from April 10 to 15 in 1980.
  - Lenient: recall 63%, precision 80%
  - Strict: recall 43%, precision 54%
McFRUMP (Mauldin 1989)

- Domain: FRUMP + Astronomy (5 scripts)
- Input: UseNet news; Output: case frame (CD)

McFRUMP (cont.)

- Vocabulary size: 3,369 nouns, 2,148 names, and 376 basic verbs + synonyms from Webster’s 7 (28,323 word senses) + near synonyms, and proper name rules.
- Interpretation Mechanism
  - PREDICTOR initiates candidate scripts.
  - SUBSTANTIATOR validates the script slots.
- Evaluation
  - Use as a component in FERRET, a conceptual information retrieval system.
  - Measured on a test corpus of 1,065 astronomy texts with recall 52% and precision 48%.

SCISOR (Rau & Jacobs 1990)

- Domain: Merger & Acquisition (M&A)
- Input: Dow Jones News + Queries
- Output: Answers
- Template: M&A (1?)
- Vocabulary size: 10,000 word roots + 75 affixes and a core concept hierarchy of 1,000 nodes.
- Interpretation Mechanism
  - TRUMPET top down processor matches bottom up output with expectations.
  - TRUMP bottom up processor parses input into frame-like semantic representation.
SCISOR (cont.)

- Evaluation:
  NA

SUMMONS (McKeown & Radev 95)

- Domain: Terrorist
- Input: Filled IE templates
- Output: Summary of a few sentences (abstract)
- Template: Terrorist (17, 25 fields)
- Vocabulary size: Large? (based on FUF, Functional Unification Formalism) (Elhadad 91, 93)
- Interpretation Mechanism
  - 7 major summary operators:
    - Change-of-perspective, contradiction, addition, refinement, agreement, superset, and trend (+ no-information)

SUMMONS (cont.)

- Evaluation:
  NA
GisTexter (Harabagiu et al. 03)

- Domain: Natural Disasters
- Input: General News
- Output: summaries (extract)
- Template: Natural Disasters (1?) + ad hoc template from WordNet topical relations.
- Vocabulary size: WordNet 2.0 + topical relations extracted from WordNet
- Interpretation Mechanism
  - Apply traditional IE to fill templates
  - Map filled template slots to text snippets
  - Generate summaries through a 5 steps incremental process

GisTexter (cont.)

- Evaluation: DUC02 multi-doc 3rd in precision and 1st in recall.

Table of contents

1. Motivation and Introduction.
2. Approaches and paradigms.
3. IE and Summarization methods.
4. Evaluating summaries.
5. The future.
Progress in Speech Recognition

MT and Summarization Evaluations

- **Machine Translation**
  - Inputs
    - Reference translation
    - Candidate translation
  - Methods
    - Manually compare two translations in:
      - Adequacy
      - Fluency
      - Informativeness
    - Auto evaluation using:
      - BLEU/NIST scores

- **Auto Summarization**
  - Inputs
    - Reference summary
    - Candidate summary
  - Methods
    - Manually compare two summaries in:
      - Content overlap
      - Linguistic qualities
    - Auto evaluation using:
      - ROUGE

MT Evaluation
Recent Development in Automatic MT Evaluation

- Problems with traditional MT evaluation
  - Expensive
  - Inconsistent
  - Long development cycle
- BLEU solution (Papineni et al. 2001)
  - Create a set of reference translations
  - Design an objective function measuring the similarity between candidate translations and reference translations
  - Calibrate the objective function with manual evaluation
  - Measure MT system performance using the objective function (Su et al. 1992)

Anatomy of BLEU Matching Score

\[
\text{BLEU} = \text{BP} \times \exp \left( \frac{1}{n} \sum_{i=1}^{n} \log \frac{p_i}{c_i} \right)
\]

- \( p_i \) is the counts of \( n \)-gram overlaps between a candidate and a reference translation.
- \( c_i \) is the total number of \( n \)-grams in the candidate translation.
- \( \text{BP} \) is weighted geometric average favors longer \( n \)-gram matches.

Improving BLEU (1)

- Brevity penalty penalizes short candidate translations by a factor of \( \exp(1-|r|/|c|) \) when \( |c| < |r| \).
- Replacing brevity penalty with standard recall measure (Melamed 2003)
- BLEU uses high order \( n \)-gram \( (n>1) \) to estimate the fluency of candidate translations but does not consider sentence level structure.
Improving BLEU (2)

- Example:
  1. police killed the gunman
  2. police kill the gunman
  3. the gunman kill police
- BLEU-2: $S_2 = S_3$ ("police", "the gunman")
- BLEU-N: uses geometric mean of unigram to N-gram precisions.
  - Any candidate translation without a N-gram match has a per-translation BLEU-N score of zero.
  - But per-translation score is desirable.

New Metric 1 – ROUGE-L

- Longest Common Subsequence (LCS)
  - Given two sequences X and Y, a longest common subsequence of X and Y is a common subsequence with maximum length.
  - Intuition
    - The longer the LCS of two translations is, the more similar the two translations are.
  - Score
    - Use LCS-based F-measure (ROUGE-L) to estimate the similarity between two translations. (see paper for more details)

ROUGE-L Example

- Example:
  1. police killed the gunman
  2. police kill the gunman
  3. the gunman kill police
- BLEU-2: $S_2 = S_3$ ("police", "the gunman")
- ROUGE-L:
  - $S_2 = 3/4$ ("police the gunman")
  - $S_3 = 2/4$ ("the gunman")
  - $S_2 > S_3$
New Metric 2 - ROUGE-W

- Weighted Longest Common Subsequence
  - Example:
    - \(X: [A B C D E F G]\)
    - \(Y_1: [A B C D H I K]\)
    - \(Y_2: [A H B K C E D]\)
    - ROUGE-L(\(Y_2\)) = ROUGE-L(\(Y_2\))
  - ROUGE-W favors strings with consecutive matches.
  - It can be computed efficiently using dynamic programming. (Lin & Och, ACL 2004)

New Metric 3 – ROUGE-S

- Skip-Bigram
  - Any pair of words in their sentence order, allowing for arbitrary gaps.
  - Intuition
    - Consider long distance dependency.
    - Allow gaps in matches as LCS but count all in-sequence pairs; while LCS only counts the longest subsequences.
  - Score
    - Use skip-bigram-based F-measure (ROUGE-S) to estimate the similarity between two translations. (see paper for more details)

ROUGE-S Example

- Example:
  1. police killed the gunman
  2. police kill the gunman
  3. the gunman kill police
  4. the gunman police killed
- BLEU-2: \(S_4>S_2=S_3\)
- ROUGE-L: \(S_2>S_3=S_4\)
- ROUGE-S:
  - \(S_2=1/6\) ("police the", "police gunman", "the gunman")
  - \(S_3=1/6\) ("the gunman")
  - \(S_4=2/6\) ("the gunman", "police killed")
  - \(S_2>S_4>S_3\)
ORANGE

- ORANGE
  - Oracle Ranking Framework for Gisting Evaluation
  - A new framework for evaluating automatic evaluation metrics
  - COLING 2004 (Lin and Och)

Other Automatic Evaluation Metrics

- Edit distance (Su et al. COLING 1992)
- RED: edit distance over multiple references (Akiba et al. 2001)
- NIST (2002): arithmetic mean of n-gram co-occurrences with information content weight based on BLEU
- WER: word error rate (Nieden et al. 2000)
- PPE: position independent WER (Leusch et al. 2003)
- GTM: generalized text matcher (Turian et al. 2003)
- ROUGE (Lin & Och ACL 2004)
  - ROUGE-L: longest common subsequence (LCS) based co-occurrences between candidate and reference translations
  - ROUGE-W: similar to ROUGE-L but favors consecutive matches
  - ROUGE-S: skip-bigram based co-occurrences between candidate and reference translations

Pearson’s ρ MT Eval 2003 Chinese

<table>
<thead>
<tr>
<th>Method</th>
<th>P 95%L</th>
<th>P 95%U</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU1</td>
<td>0.86</td>
<td>0.83</td>
</tr>
<tr>
<td>BLEU4</td>
<td>0.77</td>
<td>0.72</td>
</tr>
<tr>
<td>BLEU12</td>
<td>0.66</td>
<td>0.60</td>
</tr>
<tr>
<td>NIST</td>
<td>0.89</td>
<td>0.86</td>
</tr>
<tr>
<td>WER</td>
<td>0.47</td>
<td>0.41</td>
</tr>
<tr>
<td>PER</td>
<td>0.67</td>
<td>0.62</td>
</tr>
<tr>
<td>GTM10</td>
<td>0.82</td>
<td>0.79</td>
</tr>
<tr>
<td>GTM20</td>
<td>0.77</td>
<td>0.73</td>
</tr>
<tr>
<td>GTM30</td>
<td>0.74</td>
<td>0.70</td>
</tr>
</tbody>
</table>

*Lin & Och ACL 2004 8 systems
Problems with the Current Framework

- Comparison of automatic evaluation metrics is normally conducted at corpus level using correlation analysis between human and automatic scores. However,
  - correlation analyses across different evaluation settings are not stable
  - Human weights in factors other than adequacy and fluency
    - For example, Case vs. No Case; Stemmed vs. Not Stemmed
  - high corpus-level correlation between human and automatic metrics might not translate to high sentence-level correlation
    - For example, BLEU metric will give zero scores for all candidate translations shorter than 9 words.
  - high correlation between human and automatic scores in both adequacy and fluency cannot always be achieved
    - But an ideal metric should do well in both.

The New Evaluation Framework

- Main Idea
  - A good evaluation metric should rank a good translation higher than a bad one.
  - Reference translations are good translations.
  - Automatic translations are usually worse than references.
  - Therefore, on average a good evaluation metric should rank reference translations higher than automatic translations.
- Implementation
  - Given a NBEST list generated by a SMT system, we compute the average ranks of reference translations in the NBEST list according to different automatic evaluation metrics.
  - The higher the average ranks of reference translations are, the better the automatic evaluation metric is.

Advantages of the New Framework

- No extra human involvement
  - Use the existing human translations but not human evaluations
- Work on sentence-level
  - Diagnostic error analysis at sentence-level is naturally provided
    - a very desirable feature that is always advocated by MT researchers
- Many existing data points
  - Less prone to the data sparseness problem
    - For example, 219 sentences vs. 8 systems in MT Eval 2003 Chinese
- Only one objective function to optimize
  - Rank reference translations as high as possible
- A natural fit to existing SMT frameworks
  - For example, a metric that ranks a good sentence high in a NBEST list could be easily integrated in a minimal error rate SMT training framework (Och, ACL 2003) and improve the overall system performance.
Experiments

- MT system
  - ISI AITemp SMT system [Och, ACL 2003]
- Corpus
  - MT Eval 2003 Chinese
    - 919 sentences
    - 4 sets of reference translations
- Runs
  - BLEUS*, PER, WER, ROUGE-L (LCS), ROUGE-W (WLCS), ROUGE-S (Skip-bigram), and combinations of ROUGE-S and ROUGE-W
  - Use 1024-best list and 16384-best list
  - Estimate 95% confidence interval using bootstrap resampling
  - *smoothed version of BLEU to enable sentence level scoring

### ORANGE Scores for BLEUS1 - 9

<table>
<thead>
<tr>
<th>Method</th>
<th>ORANGE</th>
<th>Avg Rank</th>
<th>95%-CI-L</th>
<th>95%-CI-U</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEUS1</td>
<td>35.39%</td>
<td>363</td>
<td>337</td>
<td>387</td>
</tr>
<tr>
<td>BLEUS2</td>
<td>25.51%</td>
<td>261</td>
<td>239</td>
<td>283</td>
</tr>
<tr>
<td>BLEUS3</td>
<td>23.74%</td>
<td>243</td>
<td>221</td>
<td>267</td>
</tr>
<tr>
<td>BLEUS4</td>
<td>23.13%</td>
<td>237</td>
<td>215</td>
<td>258</td>
</tr>
<tr>
<td>BLEUS5</td>
<td>23.13%</td>
<td>237</td>
<td>215</td>
<td>260</td>
</tr>
<tr>
<td>BLEUS6</td>
<td>22.91%</td>
<td>235</td>
<td>211</td>
<td>257</td>
</tr>
<tr>
<td>BLEUS7</td>
<td>22.98%</td>
<td>236</td>
<td>213</td>
<td>253</td>
</tr>
<tr>
<td>BLEUS8</td>
<td>23.20%</td>
<td>238</td>
<td>214</td>
<td>261</td>
</tr>
<tr>
<td>BLEUS9</td>
<td>23.56%</td>
<td>241</td>
<td>216</td>
<td>265</td>
</tr>
</tbody>
</table>

### ORANGE Scores for ROUGE-L and ROUGE-W

<table>
<thead>
<tr>
<th>Method</th>
<th>Avg Rank Ratio</th>
<th>Avg Rank</th>
<th>Avg Rank</th>
<th>95%-CI-L</th>
<th>95%-CI-U</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROUGE-L-1</td>
<td>20.56%</td>
<td>211</td>
<td>180</td>
<td>234</td>
<td></td>
</tr>
<tr>
<td>ROUGE-W-1.1</td>
<td>20.45%</td>
<td>210</td>
<td>189</td>
<td>232</td>
<td></td>
</tr>
<tr>
<td>ROUGE-W-1.2</td>
<td>20.47%</td>
<td>210</td>
<td>186</td>
<td>230</td>
<td></td>
</tr>
<tr>
<td>ROUGE-W-1.3</td>
<td>20.69%</td>
<td>212</td>
<td>188</td>
<td>234</td>
<td></td>
</tr>
<tr>
<td>ROUGE-W-1.4</td>
<td>20.91%</td>
<td>214</td>
<td>191</td>
<td>238</td>
<td></td>
</tr>
<tr>
<td>ROUGE-W-1.5</td>
<td>21.11%</td>
<td>217</td>
<td>196</td>
<td>241</td>
<td></td>
</tr>
<tr>
<td>ROUGE-W-1.6</td>
<td>21.47%</td>
<td>220</td>
<td>199</td>
<td>242</td>
<td></td>
</tr>
<tr>
<td>ROUGE-W-1.7</td>
<td>21.73%</td>
<td>223</td>
<td>200</td>
<td>249</td>
<td></td>
</tr>
<tr>
<td>ROUGE-W-1.8</td>
<td>21.98%</td>
<td>224</td>
<td>204</td>
<td>246</td>
<td></td>
</tr>
<tr>
<td>ROUGE-W-1.9</td>
<td>22.06%</td>
<td>226</td>
<td>206</td>
<td>249</td>
<td></td>
</tr>
<tr>
<td>ROUGE-W-2.0</td>
<td>22.25%</td>
<td>228</td>
<td>208</td>
<td>250</td>
<td></td>
</tr>
</tbody>
</table>
ORANGE Scores for ROUGE-S

<table>
<thead>
<tr>
<th>Method</th>
<th>Avg Rank</th>
<th>Ratio</th>
<th>Avg Rank</th>
<th>Avg Rank 95%-CI-L</th>
<th>Avg Rank 95%-CI-U</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROUGE-S0</td>
<td>25.15%</td>
<td>258</td>
<td>224</td>
<td>261</td>
<td></td>
</tr>
<tr>
<td>ROUGE-S1</td>
<td>22.44%</td>
<td>230</td>
<td>209</td>
<td>253</td>
<td></td>
</tr>
<tr>
<td>ROUGE-S2</td>
<td>20.38%</td>
<td>209</td>
<td>195</td>
<td>231</td>
<td></td>
</tr>
<tr>
<td>ROUGE-S3</td>
<td>19.59%</td>
<td>203</td>
<td>188</td>
<td>231</td>
<td></td>
</tr>
<tr>
<td>ROUGE-S4</td>
<td>19.66%</td>
<td>202</td>
<td>177</td>
<td>224</td>
<td></td>
</tr>
<tr>
<td>ROUGE-S5</td>
<td>19.95%</td>
<td>204</td>
<td>184</td>
<td>226</td>
<td></td>
</tr>
<tr>
<td>ROUGE-S6</td>
<td>20.32%</td>
<td>208</td>
<td>187</td>
<td>231</td>
<td></td>
</tr>
<tr>
<td>ROUGE-S7</td>
<td>20.77%</td>
<td>213</td>
<td>191</td>
<td>241</td>
<td></td>
</tr>
<tr>
<td>ROUGE-S8</td>
<td>21.92%</td>
<td>222</td>
<td>204</td>
<td>242</td>
<td></td>
</tr>
<tr>
<td>ROUGE-S9</td>
<td>21.92%</td>
<td>225</td>
<td>204</td>
<td>242</td>
<td></td>
</tr>
<tr>
<td>ROUGE-S*</td>
<td>27.43%</td>
<td>281</td>
<td>259</td>
<td>304</td>
<td></td>
</tr>
</tbody>
</table>

Summary of ORANGE Scores

<table>
<thead>
<tr>
<th>Method</th>
<th>Avg Rank</th>
<th>Ratio</th>
<th>Avg Rank</th>
<th>Avg Rank 95%-CI-L</th>
<th>Avg Rank 95%-CI-U</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU8.5</td>
<td>22.91%</td>
<td>228</td>
<td>211</td>
<td>257</td>
<td></td>
</tr>
<tr>
<td>BLEU7</td>
<td>22.95%</td>
<td>228</td>
<td>211</td>
<td>257</td>
<td></td>
</tr>
<tr>
<td>PER</td>
<td>36.44%</td>
<td>373</td>
<td>350</td>
<td>403</td>
<td></td>
</tr>
<tr>
<td>NIST</td>
<td>36.44%</td>
<td>373</td>
<td>350</td>
<td>403</td>
<td></td>
</tr>
<tr>
<td>ROUGE-L</td>
<td>20.56%</td>
<td>211</td>
<td>190</td>
<td>234</td>
<td></td>
</tr>
<tr>
<td>ROUGE-W-1</td>
<td>20.45%</td>
<td>210</td>
<td>189</td>
<td>233</td>
<td></td>
</tr>
<tr>
<td>ROUGE-S4</td>
<td>19.66%</td>
<td>202</td>
<td>177</td>
<td>224</td>
<td></td>
</tr>
</tbody>
</table>

ORANGE Score vs. NBEST Cutoff
Summary of Results

- ROUGE-S4 performs the best.
  - ROUGE-S4 is skip-bigram with maximum gaps of 4 between two words.
- ROUGE-W and ROUGE-L perform significantly better than BLEUS but worst than ROUGE-S4.
- PER is the worst performer.
- BLEUS6* is the best BLEUS variant among BLEUS1 to BLEUS9.

Ongoing Work

- Verification of the superiority of reference translation assumption
  - We assumed that references are the best translations among the NBEST list, but candidate translations ranked higher than references by automatic metrics might indeed be good translations.
- Verification of the results
  - Perform discriminative training using the best automatic metric identified (i.e. ROUGE-S4) by the new framework (Och, ACL 2003).
  - Perform contrastive evaluation of the best automatic metric (ROUGE-S4) optimized translations and BLEU-4 optimized translations: ROUGE-S4 optimized results should be better.
- Developing new metrics using the new evaluation framework
  - Synonyms and paraphrases
  - Syntax-based metrics
  - Weight matches according to consensus among references

Summarization Evaluation
Document Understanding Conference (DUC)

- Part of US DARPA TIDES Project DUC 01 - 04 (http://duc.nist.gov)
  - Tasks
    - Single-doc summarization (DUC 01 and 02: 30 topics)
    - Single-doc headline generation (DUC 03: 30 topics, 04: 50 topics)
    - Multi-doc summarization
      - Generic: 50, 50, 100, 200 (2002), and 400 (2001) words summaries
      - Short summaries of about 100 words in three different tasks in 2003
      - Focused by an event (30 TDT clusters)
      - Focused by a viewpoint (50 TREC clusters)
      - In response to a question (30 TREC Novelty track clusters)
    - Short summaries of about 665 bytes in three different tasks in 2004
      - Focused by an event but documents were translated into English from Arabic: 24 topics
      - In response to a "who is X?" question (50 persons)
  - Participants
    - A new 3-year roadmap will be released later this year (DUC 2005).

Snapshot of An Evaluation Session

Overall Candidate Quality

SEE – Summary Evaluation Environment (Lin 2001)

Measuring Content Coverage

Single Reference

Coarse Judgment
Inconsistency of Human Judgment

- **Single Document Task**
  - Total of 5,921 judgments
  - Among them 1,076 (18%) contain different judgments for the same pair (from the same assessor)
  - 143 of them have three different coverage grades (2.4%)

- **Multi-Document Task**
  - Total of 6,963 judgments
  - Among them 528 (7.6%) contain different judgments for the same pair (from the same assessor)
  - 27 of them have three different coverage grades

Lin and Hovy (DUC Workshop 2002)

DUC 2003 Human vs. Human (1)

Mean Coverage Scores of Human vs. Human per Topic for DUC 2003 Task 2

DUC 2003 Human vs. Human (2)

1. Can we get consensus among humans?
2. If yes, how many humans do we need to get consensus?
3. Single reference or multiple references?
Summary of Research Issues

- How to accommodate human inconsistency?
- Can we obtain stable evaluation results despite using only a single reference summary per evaluation?
- Will inclusion of multiple summaries make evaluation more or less stable?
- How can multiple references be used in improving stability of evaluations?
- How is stability of evaluation affected by sample size?

Recent Results

- Van Halteren and Teufel (2003)
  - Stable consensus factoid summary could be obtained if 40 to 50 reference summaries were considered.
  - 50 manual summaries of one text.
  - Stable consensus semantic content unit (SCU) summary could be obtained if at least 5 reference summaries were used.
  - 10 manual multi-doc summaries for three DUC 2003 topics.
- Hori et al. (2003)
  - Using multiple references would improve evaluation stability if a metric taking into account consensus.
  - 50 utterances in Japanese TV broadcast news; each with 25 manual summaries.
  - ROUGE, an automatic summarization evaluation method used in DUC 2003.
Automatic Evaluation of Summarization Using ROUGE

- ROUGE summarization evaluation package
  - Currently (v1.5.4) include the following automatic evaluation methods: (Lin, Text Summarization Branches Out workshop 2004)
    - ROUGE-N: N-gram based co-occurrence statistics
    - ROUGE-L: LCS-based statistics
    - ROUGE-W: Weighted LCS-based statistics that favors consecutive LCSes (see ROUGE note)
    - ROUGE-S: Skip-bigram-based co-occurrence statistics
    - ROUGE-SU: Skip-bigram plus unigram-based co-occurrence statistics
  - Free download for research purpose at: http://www.isi.edu/~cyl/ROUGE

Evaluation of ROUGE

- Corpora
  - DUC 01, 02, and 03 evaluation data
  - Including human and systems summaries
- Seven task formats
  - Single doc 10 and 100 words, multi-doc 10, 50, 100, 200, and 400 words
- Three versions
  - CASE: the original summaries
  - STEM: the stemmed version of summaries
  - STOP: STEM plus removal of stopwords
- Number of references
  - Single and different numbers of multiple references
- Quality criterion
  - Pearson's product moment correlation coefficients between systems' average ROUGE scores and their human assigned mean coverage score
- Metrics
  - 17 ROUGE metrics: ROUGE-N with N = 1 to 9, ROUGE-L, ROUGE-W, ROUGE-S and ROUGE-SU (with maximum skip-distance of 0, 4, and 9)
- Statistical significance
  - 95% confidence interval estimated using bootstrap resampling

100 Words Single-Doc Task
Summary of Results

• Overall
  - Using multiple references achieved better correlation with human judgment than just using a single reference.
  - Using more samples achieved better correlation with human judgment (DUC 02 vs. other DUC data).
  - Stemming and removing stopwords improved correlation with human judgment.
  - Single-doc task had better correlation than multi-doc.

• Specific
  - ROUGE-S4, S9, and ROUGE-W1.2 were the best in 100 words single-doc task, but were statistically indistinguishable from most other ROUGE metrics.
  - ROUGE-1, ROUGE-L, ROUGE-SU4, ROUGE-SU9, and ROUGE-W1.2 worked very well in 10 words headline-like task (Pearson's ρ ~ 97%).
  - ROUGE-1, 2, and ROUGE-SU4 were the best in 100 words multi-doc task but were statistically equivalent to other ROUGE-S and SU metrics.

Effect of Sample Sizes

• Examine the effect of sample size on:
  - Human assigned mean coverage (C)
  - ROUGE
  - Correlation between C and ROUGE

• Experimental setup
  - DUC 2001 100 words single and multi-doc data.
  - Applied stemming and stopword removal.
  - Ran bootstrap resampling at sample size from 1 to 142 for single-doc task and 26 for multi-doc task.

Pearson’s ρ ROUGE-SU4 vs. Mean Coverage (DUC 2001)
Effect of Sample Size
100 Words Single-Doc Task

- DUC 2001
  - Pearson's $\rho$ ROUGE-SU4 vs. mean coverage

Effect of Sample Size
100 Words Multi-Doc Task

- DUC 2001
  - Pearson's $\rho$ ROUGE-SU4 vs. mean coverage

Summary of Results (Sample Size)

- Reliability of performance estimation in human assigned coverage score and ROUGE improved when the number of samples increased (tighter gray box).
- Reliability of correlation analysis improved when the number of samples increased:
  - Critical numbers to reach significant results in Pearson's $\rho$
    - at least 86 documents in the single-doc task.
    - at least 18 topics in the multi-doc task.
  - Number of documents and topics in DUC were well above these critical numbers.
- Using multiple reference achieved better correlation between ROUGE and mean coverage.
- Sample size had a big effect on the reliability of estimation.

Sample Size 71 76 81 86 91 96 101 126 131 136

- Sample size had a big effect on the reliability of estimation.
- Using multiple reference achieved better correlation between ROUGE and mean coverage.
- Sample size had a big effect on the reliability of estimation.
Conclusions and Future Work

- ROUGE worked pretty well in evaluations of single document summarization tasks (100 words and 10 words) with Pearson's $r$ above 85%.
- ROUGE worked fine in multi-doc summarization evaluations but how to make its metrics consistently achieve Pearson's $r$ above 85% in all test cases is still an open research topic.
- If human inconsistency is unavoidable, we then need to quantify its effect using statistical significance analysis and minimize it using large number of samples.

Basic Elements – ROUGE+

Toward Automatic Evaluation of MT and Summarization in Semantic Space

Snapshot of An Evaluation Session

Measuring Content Coverage
What Is the Right Span of Information Unit

- Information Retrieval
  - Document and passage
- Question and Answering
  - Factoid, paragraph, document, ...
- Summarization
  - Word, phrase, clause (EDU), sentence, paragraph, ...

The Factoid Method

- Factoids
  - Atomic semantic units represent sentence meaning (FOPL style).
  - "Atomic" means that a semantic unit is used as a whole across multiple summaries.
  - Each factoid may carry information varying from a single word to a clause.
- Example:
  - The police have arrested a white Dutch man.
    - A suspect was arrested.
    - The police did the arresting.
    - The suspect is white.
    - The suspect is Dutch.
    - The suspect is male.

The Pyramid Method

- Pyramid
  - A weighted inventory of factoids or summarization content units (SCU)
    - A: "Unable to make payments on a $2.1 billion debt"
    - B: "made payments on PAL's $2 billion debt impossible"
    - C: "with a rising $2.1 billion debt"
    - D: "PAL is buried under a $2.2 billion dollar debt it cannot repay"
    - SCU
      - F1: PAL has 2.1 million debt (All)
      - F2: PAL can't make payments on debt (Most)
Problems with Factoid and SCU

- Each factoid may carry very different amount of information
  - How to assign fair information value to a factoid?
- The inventory of factoids grows as more summaries are added to the reference pool
  - Old factoids break apart to create new factoids
- Interdependency of factoids are ignored
- Totally manual creation so far and only been tested on very small data set
  - Factoid: 2 documents
  - SCU+Pyramid: 3 sets of multi-doc topics

How to automate?

Basic Elements (BE)

- Definition
  - A head, modifier and relation triple: BE::<HEAD|MOD|REL>
    - BE::HEAD is the head of a major syntactic constituent (noun, verb, adjective or adverbial phrases).
    - BE::MOD is a single dependent of BE::HEAD with a relation, BE::REL, between them.
    - BE::REL could be a syntactic, semantic relation or NIL.
- Example
  - “Two Libyans were indicted for the Lockerbie bombing in 1991.”
    - <$\text{Libyans} | \text{two} | \text{CARDINAL}$>
    - <$\text{indicted} | \text{Libyans} | \text{ACCUSED}$>
    - <$\text{indicted} | \text{bombing} | \text{CRIME}$>
    - <$\text{indicted} | \text{1991} | \text{TIME}$>

Research Issues

- How can BEs be created automatically?
  - Extract dependency triples from automatic parse trees.
    - BE-F: MINPAP triples* (Lin 95)
    - BE-L: Charniak parse trees + automatic semantic role tagging*
- What score should each BE have?
  - Equal weight*, tfidf, information value, ...
- When do two BEs match?
  - Lexical*, lemma*, synonym, distributional similarity, ...
- How should an overall summary score be derived from the individual matched BEs' scores?
  - Consensus of references*
Current Status

- First version, BE 1.0, released to the research community on April 13, 2005.
  - Package include:
    - BE-F (Minipar) BE breakers
    - ROUGE-1.5.4 scorer
  - One of the three official automatic evaluation metrics for Multilingual Summarization Evaluation 2005 (MSE 2005).
  - It will also be used in DUC 2005.
  - Free download for research purpose at: http://www.isi.edu/~cyl/BE

Evaluation

- Measurement
  - Examine the Pearson’s correlation between human assigned mean coverage (C) and BE.
  - Compare results with ROUGE 1-4, S4, and SU4.
- Experimental setup
  - Use DUC 2002 (10 systems) and 2003 (18 systems) 100 words multi doc data.
  - Compare single vs. multiple references.
  - Applied stemming and stopword removal.

Correlation Analysis (DUC 2002)
Correlation Analysis (DUC 2003)

<table>
<thead>
<tr>
<th>Original</th>
<th>Stopped</th>
<th>Original</th>
<th>Stopped</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>0.784</td>
<td>0.776</td>
<td>0.785</td>
</tr>
<tr>
<td>HM</td>
<td>0.959</td>
<td>0.949</td>
<td>0.917</td>
</tr>
<tr>
<td>HMR</td>
<td>0.882</td>
<td>0.864</td>
<td>0.753</td>
</tr>
<tr>
<td>HM1</td>
<td>0.859</td>
<td>0.847</td>
<td>0.853</td>
</tr>
<tr>
<td>HMR1</td>
<td>0.961</td>
<td>0.952</td>
<td>0.921</td>
</tr>
<tr>
<td>HMR2</td>
<td>0.860</td>
<td>0.848</td>
<td>0.855</td>
</tr>
</tbody>
</table>

Oracle Experiment

- We would like to answer the following questions:
  1. How well the topic identification module of a summarization system work?
  2. How much information do we lose by only matching on lexical level?
  3. How much do humans agree to each other?

- Experimental setup:
  - Use DUC 2003 multi-doc data with 4 human refs.
  - Use BE ranking to simulate topic identification. We compare three BE ranking methods:
    - LR (log likelihood ratio), DF (document frequency with LR), and RawDF (raw document frequency)
  - Use BE-F as counting unit.
  - Compute recall and precision of different BE ranking methods vs. reference summary BE ranking based on raw DF.

Oracle Exp: BE-F Recall
Conclusions

- BE-F consistently achieves over 90% Pearson’s correlation with human judgments in all testing categories.
  - BE-F with stemming and matching only on BE::HEAD and BE::MOD (HH & HHH) has the best correlation.
- BE-L has over 90% correlation when both BE::HEAD and BE::MOD are considered in the matching. It also works better with multiple references.
- BE-F and BE-L are more stable than ROUGE across corpora (DUC’02 R2 Org vs. DUC’03 R3 Stop)
- Need to go beyond lexical matching.
- Need to develop better BE ranking algorithms.
- Need to address the issue of human disagreement:
  - Better summary writers?
  - Better domain knowledge?
  - Better task definition …

Future Directions

- BE breaking
  - Use FrameNet II frame elements in BE relations.
- BE matching
  - Paraphrases, synonyms, and distributional similarity.
- BE ranking
  - Prioritize BEs in a given application context.
  - Assign weights according to BE’s information content.
  - Utilize inter-BE dependency.
- Application
  - Develop summarization methods based on BE.
Table of contents

1. Motivation and Introduction.
2. Approaches and paradigms.
3. IE and Summarization methods.
4. Evaluating summaries.
5. The future.

What's Next?

- Robust information extraction on general domains based on: FrameNet, PropBank, WordNet and other knowledge sources.
- Automatic frame acquisition and integration to increase coverage.
- Develop better template/frame merging and linking method.
- Construct large common summarization corpora and design better evaluation methods to facilitate more accurate fast develop and test cycles.

再会 Good Bye!
References (4)


Online bibliographies:
- http://www.isi.edu/~cyl/summarization
- http://www.dcs.shef.ac.uk/~gael/alphalist.html