Evaluation and Applications of Automatic Text Summarization

Chin-Yew LIN
Natural Language Group
Information Sciences Institute
University of Southern California
cyl@isi.edu
http://www.isi.edu/~cyl

Automatic Speech Recognition - A Success Story


Chin-Yew LIN, NTCIR-5, Tokyo, Japan, Dec 9, 2005
**Statistical Machine Translation - Another Success Story?**

- Goal: Automatic translation of texts from one natural language to another
- Common components of statistical machine translation (SMT) systems
  - *Translation model, language model, decoder, and evaluator*
  
```
<table>
<thead>
<tr>
<th>Foreign/Japanese Parallel Text</th>
<th>Japanese Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical Analysis</td>
<td>Statistical Analysis</td>
</tr>
<tr>
<td>Translation Model</td>
<td>Language Model</td>
</tr>
<tr>
<td>Decoder</td>
<td>Automatic Evaluator</td>
</tr>
<tr>
<td>Japanese Reference Text</td>
<td>Translation Quality Estimation</td>
</tr>
</tbody>
</table>
```

**Document Understanding Conference (DUC)**

- **Tasks (DUC 2001 - 2006, NIST, USA)**
  - Single-doc summarization (DUC 01 and 02: 30 topics)
  - Single-doc headline generation (DUC 03: 30 topics, 04: 50 topics)
  - Multi-doc summarization
    - Generic 10, 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 (30 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 (50 TDT clusters)
      - focused by an event but documents were translated into English from Arabic (24 topics)
      - in response to a “who is X?” question (50 persons)
  - **DUC 2005 and 2006 (50 topics): Question-focused summarization task.**
    - Real-world complex question answering, in which an information need cannot be satisfied by simply stating a name, date, quantity, etc. Given a question and a set of 25 relevant documents, the task is to synthesize a fluent, well-organized 250-word summary of the documents that answers the question(s) in the topic statement.

- **Participants**
Summary of 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.

- **Nenkova and Passonneau (2003)**
  - 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.

- **Lin and Hovy (2003), Lin (2004)**

- **Hovy, Lin, Zhou, and Fukumoto (2005)**
  - Basic elements (BE), a new automatic summarization evaluation method intending to move beyond simple surface level word/stem matching and into semantic matching. BE has been used DUC 2005 and showed good correlation with human judgments. (http://www.summaries.net/BE)
An Information-Theoretic Approach to Automatic Evaluation of Summaries

incorporation with
Guihong Cao*, Jienfeng Gao#, and
Jian-Yun Nie*
*University of Montreal
#Microsoft Corporation

Summarization as a Generative Process

- Given a set of documents \( D = \{d_1, d_2, \ldots, d_i\} \)
  - \( i = 1 \) for single document summarization
  - \( i > 1 \) for multiple document summarization,
- We assume there exists a probabilistic distribution with parameters specified by \( \theta \) that generates a summary \( S \) from \( D \).
- The task of automatic summarization is to estimate \( \theta \) that maximizes the likelihood of a set of target summaries \( S^{1:L} \) given a set of input document sets \( D^{1:L} \):
  \[
  \hat{\theta} = \arg \max_{\theta} p(S^{1:L} | \theta, D^{1:L})
  \]
Summarization Evaluation — an Information-Theoretic View

• Given $\theta_R$ that generates reference summaries and $\theta_A$ that generates system summaries:
  • A better system summary should have a better $\theta_A$ that is close to $\theta_R$.
  • The task of summarization evaluation is to estimate the distance between $\theta_A$ and $\theta_R$.
  • Possible distance measures:
    • Kullback-Leibler divergence (KL)
    • Jensen-Shannon divergence (JS)
  • We propose:

$$
\text{Score}_{\text{summary}}^{JSD}(S_A | S_R^{1:L}) = -JS_{1/2}(p(\theta_A | S_A) \parallel p(\theta_R | S_R^{1:L}))
$$

Kullback-Leibler vs. Jensen-Shannon

• Kullback-Leibler Divergence

$$
KL(p_1 \parallel p_2) = \sum_{\theta} p_{\theta} \log \left( \frac{p_{\theta}}{p_2} \right)
$$

• KL divergence has discontinuity over its sampling space; it’s undefined where $p_2=0$.
• KL divergence is asymmetric, i.e.

$$
KL(p_1 \parallel p_2) \neq KL(p_2 \parallel p_1)
$$

• Jensen-Shannon Divergence (Lin 1991)

$$
JS_{1/2}(p_1 \parallel p_2) = \frac{1}{2} \sum_{\theta} p_{\theta} \log \left( \frac{p_{\theta}}{\frac{1}{2} p_1 + \frac{1}{2} p_2} \right) + p_{\theta} \log \left( \frac{p_{\theta}}{\frac{1}{2} p_1 + \frac{1}{2} p_2} \right)
$$
How to Estimate $\theta$?

• Assume a multinomial summary generation model (Zaragoza et al. 2003):
  \[ \theta = (\theta_1, \theta_2, \ldots, \theta_v) \in [0,1]^v, \quad \sum_{i=1}^v \theta_i = 1 \]

• Instead of estimating $\theta$, we estimate its posterior using Bayes’ rule:
  \[ p(\theta | S) = \frac{p(S | \theta) p(\theta)}{p(S)} \]

• Assuming a multinomial unigram model and by choosing a Dirichlet prior for $p(\theta)$, we have the posterior probability also in Dirichlet form that has a maximum a posterior (MAP) estimation as follows (Gelman et al. 2003):
  \[ \theta_{MAP} = \frac{C(w_i,S) + \alpha_i - 1}{\sum_{i=1}^V (C(w_i,S) + \alpha_i) - V} \]

  Hyperparameter for Dirichlet prior

Multinomial Distribution and Dirichlet Prior

• Multinomial distribution
  \[ p(S | \theta) = Z_{a_0} \prod_{i=1}^V (\theta_i)^{a_i}; \quad a_i = C(w_i | S) \]
  \[ a_0 = \sum_{i=1}^V a_i; \quad Z_{a_0} = \frac{\Gamma(a_0 + 1)}{\prod_{i=1}^V \Gamma(a_i + 1)} \]

• Dirichlet Prior
  \[ p(\theta) = Z'_{a_0} \prod_{i=1}^V (\theta_i)^{\alpha_i - 1}; \quad \alpha_i \geq 1 \]
  \[ \alpha_0 = \sum_{i=1}^V \alpha_i; \quad Z'_{a_0} = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^V \Gamma(\alpha_i + 1)} \]
### Smoothing $\theta$

\[
\theta_i^{MAP} = \frac{C(w_i, S) + \alpha_i - 1}{\sum_{j=1}^{V} (C(w_j, S) + \alpha_j) - V}
\]
\[
\theta_i^{ML} = \frac{C(w_i, S)}{\sum_{j=1}^{V} C(w_j, S)}; \quad \alpha_i = 1
\]
\[
\theta_i^{Additive} = \frac{C(w_i, S) + \lambda}{\sum_{j=1}^{V} C(w_j, S) + \lambda V}; \quad \alpha_i = \lambda + 1, \lambda > 0
\]
\[
\theta_i^{Bayes} = \frac{C(w_i, S) + \mu p(w_i | T)}{\sum_{j=1}^{V} C(w_j, S) + \mu}; \quad \alpha_i = \mu p(w_i | T) + 1
\]

### Evaluation

- **Measurement**
  - Examine the Pearson’s and Spearman’s correlations between human assigned mean coverage and automatic scores:
    - Jensen-Shannon divergence without smoothing (JSD)
    - Jensen-Shannon divergence with Bayes-smoothing (JSDS)
    - Kullback-Leibler divergence with Bayes-smoothing (KLDS)
    - Log likelihood ratio with Bayes-smoothing (LLS)

  \[
  Score_{summary}^{LLS} (S_A | S_R^{1:L}) = \sum_{i=1}^{N_A} \log p(\theta_i^{Bayes} | S_R^{1:L})
  \]

- **Experimental setup**
  - Use DUC 2002 100 words single and multi doc data.
  - Compare single vs. multiple references.
  - Apply stemming but keep stopwords.
  - Set Bayes-smoothing factor $\mu$ to 2000. (Zhai & Lafferty 04)
Conclusions & Future Directions

- Information-theoretic measure based on Jensen-Shannon divergence ($JSD$) without smoothing performed the best among all measures.
- $JSD$-based measure also compared favorably to unigram-based ROUGE-1, especially in the multi-document summarization task.
- $JSD$-based measure did as well as ROUGE based on longer N-grams. We would like to extend our unigram-based bag-of-words multinomial generation model into N-gram-based bag-of-N-grams multinomial generation model.
- Smoothed measures did not do well. This is not a surprise due to the nature of the task of summarization evaluation. Intuitively, only information presented in system summaries could be accounted for scoring:
  - What are in reference summaries should also be in good system summaries;
  - System summaries should not be given credit for information they do not provide.
- $JSD$-based measure still match only on lexical level $\Rightarrow$ apply query expansion technique to move toward matching in semantic space.
  - Use Markov chain expansion proposed by Lafferty & Zhai (2001)
  - Use information-flow expansion proposed by Nie & Cao (2005)
  - Use probabilistic latent semantic analysis (PLSA) proposed by Hoffmann (1999)
Summarization Applications

Browse a Summarized Web

- The Palm m100 handheld is the first product in the new Entry Level Product Line, where it is positioned as the entry-level consumer Palm product.

* Stanford PowerBrowser Project, Orkut et al. WWW10, 2001
**Summarizing Public Opinions and Press Coverage**

As Bush starts 2nd term, wild election media turned mild
Rating of Bush: network news, Fox, Time, Newsweek, 1st quarter ’02, ’03, ’04, ’05

Although Bush is far from the high standing in the media that he had after 9-11, in the first quarter of 2003, he started with a remarkably mild media watching him go into his 2nd term. One year ago at the kick-off of the election campaign, he faced harsh media criticism.

(MediaTenor: http://www.mediatenor.com)

Chin-Yew LIN, NTCIR-5, Tokyo, Japan, Dec 9, 2005

---

**Summarizing Product Reviews**

Welcome to Opinion Observer


Chin-Yew LIN, NTCIR-5, Tokyo, Japan, Dec 9, 2005
Summarizing Research Trend (Lee et al. CHI 2005)

ISI – DARPA Surprise Language Exercise 2003
(Leuski et al. 03)
Thank You!

Automatic Text Summarization - Another Success Story?

- Goal: Automatic translation of texts from one natural language to another
- Common components of statistical machine translation (SMT) systems
  - Translation model, language model, decoder, and evaluator
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, ...

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 (v1.5.5) 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

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?
  - No predetermined size of factoids or SCUs ⇒ “counting matches” and “scoring” would be problematic.
- The inventory of factoids grows as more summaries are added to the reference pool
  - Old factoids tend to 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.”
    - ⇒ <Libyans|two|CARDINAL>
    - ⇒ <indicted|Libyans|ACCUSED>
    - ⇒ <indicted|bombing|CRIME>
    - ⇒ <indicted|1991|TIME>

Research Issues

- **How can BEs be created automatically?**
  - Extract dependency triples from automatic parse trees.
    - BE-F: MINPAR 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-1.5.5 scorer
  - One of the three official automatic evaluation metrics for Multilingual Summarization Evaluation 2005 (MSE 2005).
  - It is 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)

#### DUC-2002 M100 BE-F vs. Human Scores Pearson's Correlation

<table>
<thead>
<tr>
<th>Original</th>
<th>Stemmed</th>
<th>Stopped</th>
<th>Original</th>
<th>Stemmed</th>
<th>Stopped</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>0.751</td>
<td>0.755</td>
<td>0.837</td>
<td>0.698</td>
<td>0.707</td>
</tr>
<tr>
<td>RM1</td>
<td>0.619</td>
<td>0.609</td>
<td>0.773</td>
<td>0.622</td>
<td>0.611</td>
</tr>
<tr>
<td>RM2</td>
<td>0.736</td>
<td>0.733</td>
<td>0.647</td>
<td>0.488</td>
<td>0.501</td>
</tr>
<tr>
<td>RL</td>
<td>0.647</td>
<td>0.633</td>
<td>0.667</td>
<td>0.679</td>
<td>0.726</td>
</tr>
<tr>
<td>RS4</td>
<td>0.747</td>
<td>0.748</td>
<td>0.845</td>
<td>0.723</td>
<td>0.726</td>
</tr>
<tr>
<td>RSU4</td>
<td>0.855</td>
<td>0.865</td>
<td>0.867</td>
<td>0.809</td>
<td>0.822</td>
</tr>
</tbody>
</table>

#### DUC-2002 M100 BE-L vs. Human Scores Pearson's Correlation

<table>
<thead>
<tr>
<th>Original</th>
<th>Stemmed</th>
<th>Stopped</th>
<th>Original</th>
<th>Stemmed</th>
<th>Stopped</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>0.933</td>
<td>0.927</td>
<td>0.912</td>
<td>0.896</td>
<td>0.889</td>
</tr>
<tr>
<td>RM1</td>
<td>0.895</td>
<td>0.880</td>
<td>0.857</td>
<td>0.857</td>
<td>0.867</td>
</tr>
<tr>
<td>RM2</td>
<td>0.719</td>
<td>0.717</td>
<td>0.837</td>
<td>0.667</td>
<td>0.667</td>
</tr>
<tr>
<td>RL</td>
<td>0.811</td>
<td>0.817</td>
<td>0.886</td>
<td>0.504</td>
<td>0.519</td>
</tr>
<tr>
<td>RS4</td>
<td>0.961</td>
<td>0.952</td>
<td>0.921</td>
<td>0.921</td>
<td>0.914</td>
</tr>
<tr>
<td>RSU4</td>
<td>0.865</td>
<td>0.865</td>
<td>0.867</td>
<td>0.809</td>
<td>0.822</td>
</tr>
</tbody>
</table>

### Correlation Analysis (DUC 2003)

#### DUC-2003 M100 BE-F vs. Human Scores Pearson's Correlation

<table>
<thead>
<tr>
<th>Original</th>
<th>Stemmed</th>
<th>Stopped</th>
<th>Original</th>
<th>Stemmed</th>
<th>Stopped</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>0.931</td>
<td>0.927</td>
<td>0.920</td>
<td>0.940</td>
<td>0.940</td>
</tr>
<tr>
<td>RM1</td>
<td>0.933</td>
<td>0.933</td>
<td>0.904</td>
<td>0.930</td>
<td>0.930</td>
</tr>
<tr>
<td>RM2</td>
<td>0.931</td>
<td>0.931</td>
<td>0.904</td>
<td>0.930</td>
<td>0.930</td>
</tr>
<tr>
<td>RL</td>
<td>0.933</td>
<td>0.933</td>
<td>0.933</td>
<td>0.933</td>
<td>0.933</td>
</tr>
<tr>
<td>RS4</td>
<td>0.895</td>
<td>0.895</td>
<td>0.895</td>
<td>0.895</td>
<td>0.895</td>
</tr>
<tr>
<td>RSU4</td>
<td>0.859</td>
<td>0.859</td>
<td>0.859</td>
<td>0.859</td>
<td>0.859</td>
</tr>
</tbody>
</table>

#### DUC-2003 M100 BE-L vs. Human Scores Pearson's Correlation

<table>
<thead>
<tr>
<th>Original</th>
<th>Stemmed</th>
<th>Stopped</th>
<th>Original</th>
<th>Stemmed</th>
<th>Stopped</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>0.959</td>
<td>0.949</td>
<td>0.917</td>
<td>0.918</td>
<td>0.918</td>
</tr>
<tr>
<td>RM1</td>
<td>0.882</td>
<td>0.882</td>
<td>0.753</td>
<td>0.718</td>
<td>0.718</td>
</tr>
<tr>
<td>RM2</td>
<td>0.859</td>
<td>0.859</td>
<td>0.853</td>
<td>0.849</td>
<td>0.849</td>
</tr>
<tr>
<td>RL</td>
<td>0.961</td>
<td>0.952</td>
<td>0.921</td>
<td>0.914</td>
<td>0.914</td>
</tr>
<tr>
<td>RS4</td>
<td>0.811</td>
<td>0.817</td>
<td>0.744</td>
<td>0.754</td>
<td>0.754</td>
</tr>
<tr>
<td>RSU4</td>
<td>0.747</td>
<td>0.748</td>
<td>0.845</td>
<td>0.723</td>
<td>0.726</td>
</tr>
</tbody>
</table>

Chin-Yew LIN, NTCIR-5, Tokyo, Japan, Dec 9, 2005
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 (HM & HM1) 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.