USC

USC Viterbi School of Engineering

ISI Information Sciences Institute

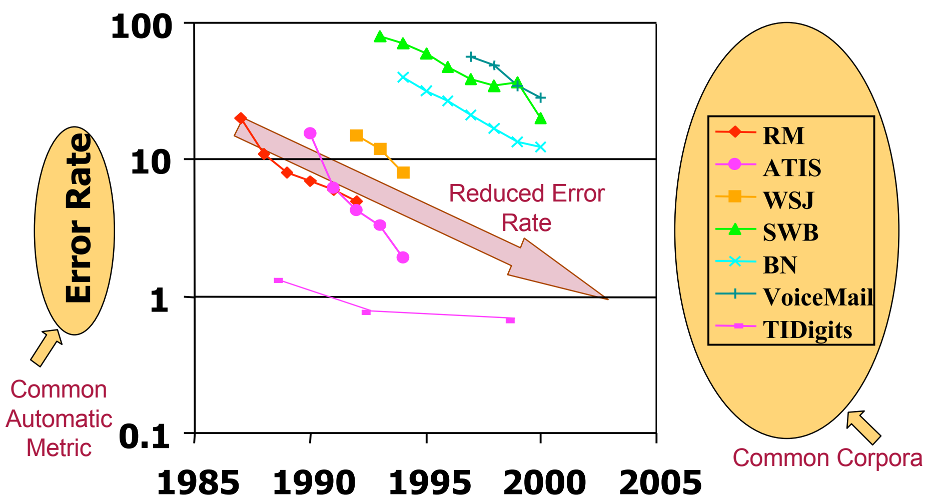
Evaluation and Applications of Automatic Text Summarization

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

USC Viterbi School of Engineering



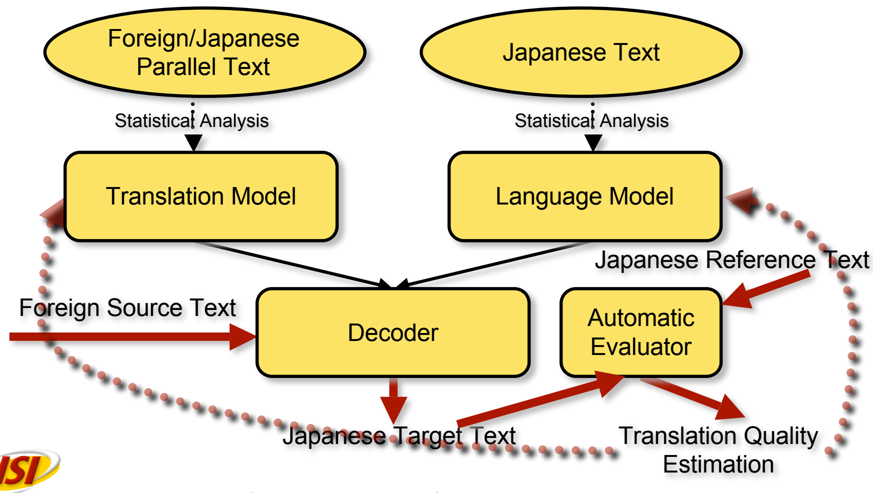
Mukund Padmanabhan and Michael Picheny, HLT Group, IBM Research, ASR Workshop, Paris, Sep 2000.

ISI Information Sciences Institute

Statistical Machine Translation - Another Success Story?

USC Viterbi School of Engineering

- 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***



ISI Information Sciences Institute

Document Understanding Conference (DUC)

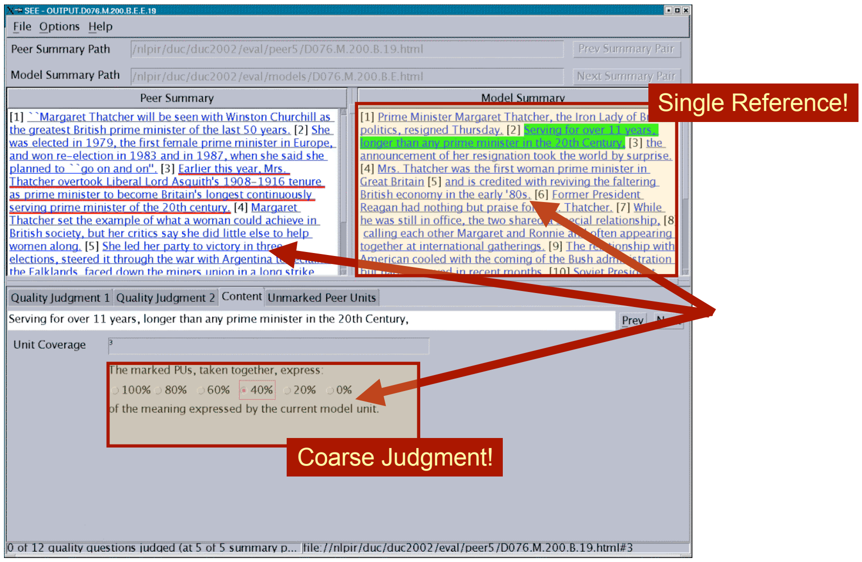
USC Viterbi School of Engineering

- 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
    - 15 systems in DUC 2001, 17 in DUC 2002, 21 in DUC 2003, 25 in DUC 2004, and 31 in DUC 2005.

ISI Information Sciences Institute

Snapshot of an Evaluation Session Measuring Content Coverage

USC Viterbi School of Engineering



Summary of Recent Results

USC Viterbi School of Engineering

- 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)
  - ROUGE, an automatic summarization evaluation method used in DUC 2003, 2004, and 2005. ROUGE is the current de facto automatic evaluation method in text summarization. (http://www.summaries.net/ROUGE)
- 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)

USC

USC Viterbi School of Engineering

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

USC Viterbi School of Engineering

- 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 θ that generates a summary S from D.
- The task of automatic summarization is to estimate θ that maximizes the likelihood of a set of target summaries $S^{1,L}$ given a set of input document sets $D^{1,L}$:

$$\hat{\theta} = \arg\max_{\theta} p(S^{1,L} \mid \theta, D^{1,L})$$

Summarization Evaluation — an Information-Theoretic View

USC Viterbi School of Engineering

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

$$Score_{summary}^{JSD}(S_A \mid S_R^{1,L}) = -JS_{1/2}(p(\theta_A \mid S_A) \,\|\, p(\theta_R \mid S_R^{1,L}))$$

ISI Information Sciences Institute

Kullback-Leibler vs. Jensen-Shannon

USC Viterbi School of Engineering

- Kullback-Leibler Divergence

$$KL(p_1 \,\|\, p_2) = \sum_{\theta_i}\left(p_1 \log\left(\frac{p_1}{p_2}\right)\right)$$

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

$$KL(p_1 \,\|\, p_2) \neq KL(p_2 \,\|\, p_1)$$

- Jensen-Shannon Divergence (Lin 1991)

$$JS_{1/2}(p_1 \,\|\, p_2) = \frac{1}{2}\sum_{\theta_i}\left(p_1 \log\left(\frac{p_1}{\frac{1}{2}p_1 + \frac{1}{2}p_2}\right) + p_2 \log\left(\frac{p_2}{\frac{1}{2}p_1 + \frac{1}{2}p_2}\right)\right)$$

ISI Information Sciences Institute

How to Estimate θ ?

- 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 θ, we estimate its posterior using Bayes' rule:

$$p(\theta \mid S) = \frac{\underbrace{\boxed{p(S \mid \theta)}}_{\text{summary likelihood}}\;\overbrace{\boxed{p(\theta)}}^{\text{prior}}}{p(S)}$$

  - 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):

$$\theta_i^{MAP} = \frac{C(w_i, S) + \alpha_i - 1}{\sum_{i=1}^{V} \left(C(w_i, S) + \alpha_i \right) - V}$$

Hyperparameter for Dirichlet prior

Multinomial Distribution and Dirichlet Prior

- Multinomial distribution

$$p(S \mid \theta) = Z_{a_0} \prod_{i=1}^{V} (\theta_i)^{a_i}; \quad a_i = C(w_i \mid S)$$

$$a_0 = \sum_{i=1}^{V} a_i; \quad Z_{a_0} = \frac{\Gamma(a_0 + 1)}{\prod_{i=1}^{V} \Gamma(a_i + 1)}$$

- Dirichlet Prior

$$p(\theta) = Z'_{\alpha_0} \prod_{i=1}^{V} (\theta_i)^{\alpha_i - 1}; \quad \alpha_i \geq 1$$

$$\alpha_0 = \sum_{i=1}^{V} \alpha_i; \quad Z'_{\alpha_0} = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{V} \Gamma(\alpha_i + 1)}$$

Smoothing θ

$$\theta_i^{MAP} = \frac{C(w_i,S)+\alpha_i - 1}{\sum_{i=1}^{V}\left(C(w_i,S)+\alpha_i\right) - V}$$

$$\theta_i^{ML} = \frac{C(w_i,S)}{\sum_{i=1}^{V} C(w_i,S)}; \quad \alpha_i = 1$$

$$\theta_i^{Additive} = \frac{C(w_i,S)+\lambda}{\sum_{i=1}^{V} C(w_i,S)+\lambda V}; \quad \alpha_i = \lambda + 1, \lambda > 0$$

$$\theta_i^{Bayes} = \frac{C(w_i,S)+\mu p(w_i \mid T)}{\sum_{i=1}^{V} C(w_i,S)+\mu}; \quad \alpha_i = \mu p(w_i \mid T) + 1$$

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)

$$Score_{summary}^{LLS}(S_A \mid S_R^{1,L}) = \sum_{i=1}^{|S_A|} \log p(\theta_i^{Bayes} \mid S_R^{1,L})$$

- 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 μ to 2000. (Zhai & Lafferty 04)

Correlation Analysis (DUC 2002)

USC Viterbi School of Engineering

		JSD		JSDS		KLDS		LLS	
		P	S	P	S	P	S	P	S
Single-Doc	Single-Ref	**0.967**	**0.911**	0.612	0.246	0.594	0.233	-0.544	0.158
	Multi-Ref	**0.969**	**0.911**	0.620	0.646	0.610	0.246	-0.599	0.114
Multi-Doc	Single-Ref	**0.803**	**0.830**	0.439	0.636	0.343	0.539	0.215	0.358
	Multi-Ref	**0.881**	**0.891**	0.761	0.806	0.606	0.709	0.474	0.600

	ROUGE-1		ROUGE-2		ROUGE-3		ROUGE-4	
	P	S	P	S	P	S	P	S
Single-Doc	0.986	0.836	0.998	0.961	0.997	0.981	0.996	0.990
Multi-Doc	0.701	0.588	0.890	0.842	0.922	0.854	0.901	0.782

ISI Information Sciences Institute

Conclusions & Future Directions

USC Viterbi School of Engineering

- 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 ⇒ 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)

ISI Information Sciences Institute

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

USC

USC Viterbi
School of Engineering

ISI
Information Sciences Institute

Summarization Applications

Browse a Summarized Web

USC Viterbi
School of Engineering

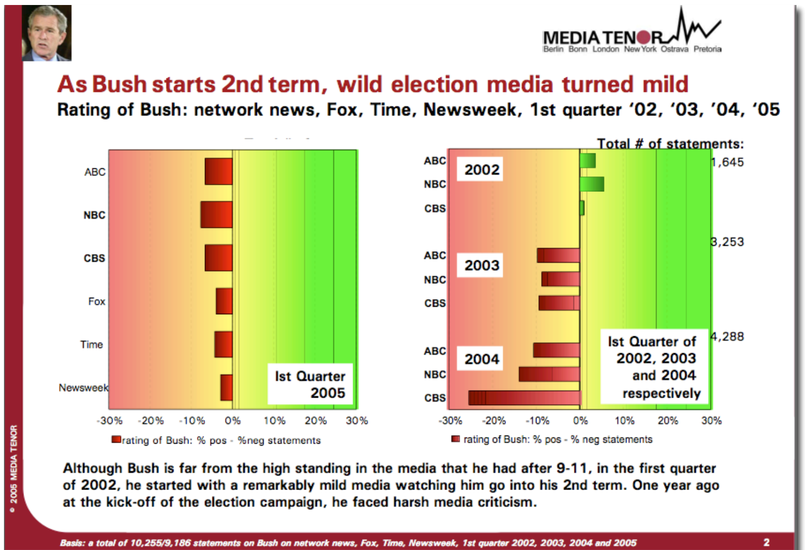
- The Palm m100 handheld is the f
- The Palm m100 handheld is the first product in the new Entry Level Product Line, where it is
- The Palm m100 handheld is the first product in the new Entry Level Product Line, where it is positioned as the entry-level consumer Palm product.

* Stanford PowerBrowser Project, Orkut et al. WWW10, 2001

ISI
Information Sciences Institute

Summarizing Public Opinions and Press Coverage

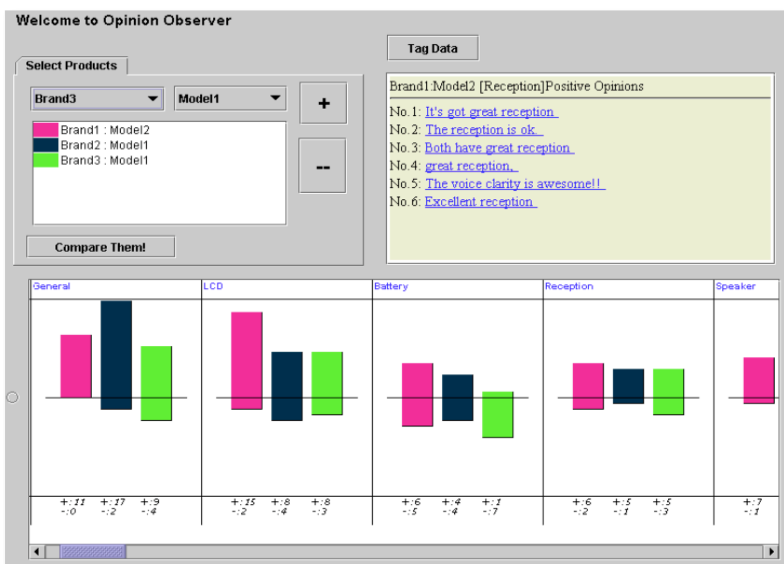
USC Viterbi School of Engineering



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

Summarizing Product Reviews

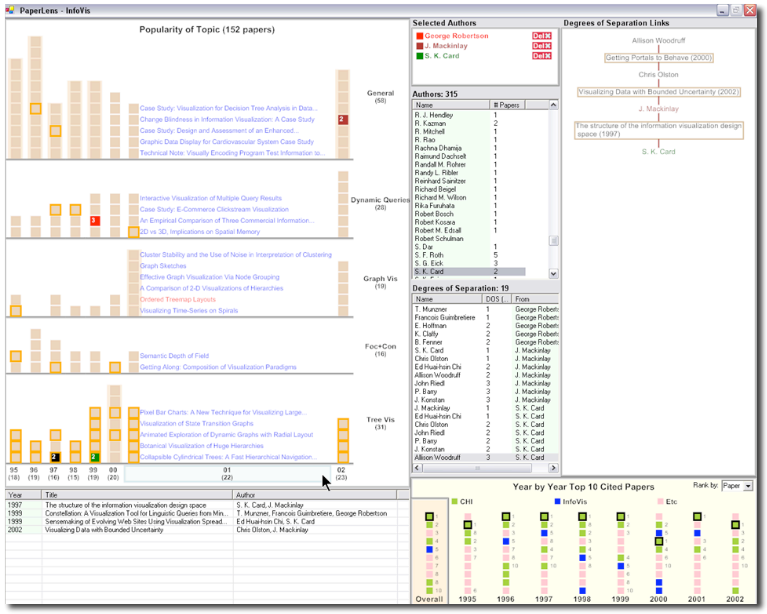
USC Viterbi School of Engineering



(Bing Liu et al., "Opinion Observer: Analyzing and Comparing Opinions on the Web", WWW 2005)

Summarizing Research Trend (Lee et al. CHI 2005)

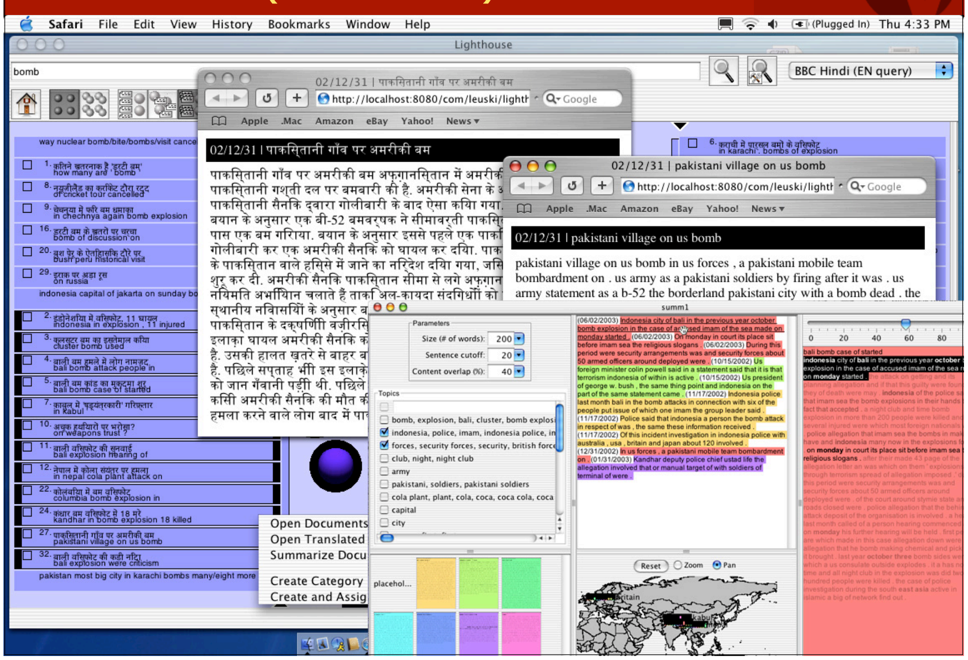
USC Viterbi School of Engineering



ISI Information Sciences Institute

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

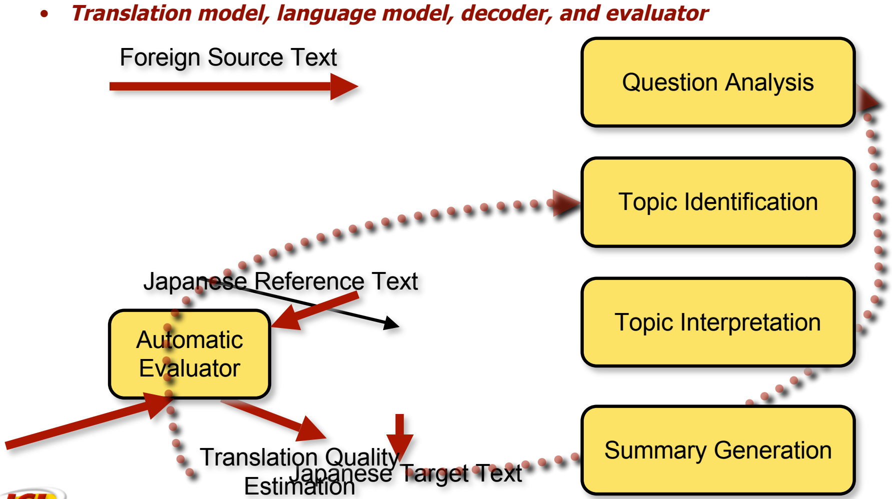
USC Viterbi School of Engineering



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.
- 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)
  - ROUGE, an automatic summarization evaluation method used in DUC 2003.

Automatic Evaluation of Summarization Using ROUGE

USC Viterbi School of Engineering

- 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

ISI Information Sciences Institute

The Factoid Method

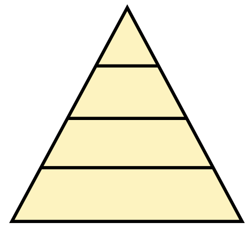
USC Viterbi School of Engineering

- Van Halteren & Teufel (2003, 2004)
- 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.

ISI Information Sciences Institute

The Pyramid Method

USC Viterbi School of Engineering



- Nenkova and Passonneau (2003)
- 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)

ISI Information Sciences Institute

Problems with Factoid and SCU

USC Viterbi School of Engineering

- 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?**

ISI Information Sciences Institute

Basic Elements (BE)

USC Viterbi School of Engineering

- 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 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>

ISI Information Sciences Institute

Research Issues

USC Viterbi School of Engineering

- 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*

ISI Information Sciences Institute

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

Current Status

USC Viterbi School of Engineering

- 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

Information Sciences Institute

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

Evaluation

USC Viterbi School of Engineering

- 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.

Information Sciences Institute

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

Correlation Analysis (DUC 2002)

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

	multi-ref		single-ref	
	Original	Stemmed	Original	Stemmed
H	NA	NA	NA	NA
HM	0.914	0.915	0.924	**0.953**
HMR	0.909	0.907	0.934	**0.953**
HM1	0.914	0.915	0.924	**0.953**
HMR1	0.909	0.907	0.934	**0.953**
HMR2	0.909	0.907	0.934	**0.953**

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

	multi-ref		single-ref	
	Original	Stemmed	Original	Stemmed
H	0.890	0.880	0.874	0.871
HM	0.917	0.932	0.865	0.895
HMR	0.917	**0.951**	0.815	0.894
HM1	0.907	0.902	0.879	0.887
HMR1	0.921	0.932	0.867	0.881
HMR2	0.909	0.904	0.879	0.882

DUC-2002 ROUGE vs. Human Scores Pearson's Correlation

	multi-ref			single-ref		
	Original	Stemmed	Stopped	Original	Stemmed	Stopped
R1	0.751	0.755	0.837	0.698	0.707	0.835
R2	0.933	0.927	0.912	0.896	0.889	0.873
R3	**0.962**	0.959	0.914	0.931	0.922	0.855
R4	0.924	0.918	0.889	0.911	0.901	0.773
RL	0.719	0.717	0.837	0.667	0.667	0.820
RS4	0.895	0.906	0.881	0.857	0.867	0.860
RSU4	0.855	0.865	0.867	0.809	0.822	0.853

Correlation Analysis (DUC 2003)

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

	multi-ref		single-ref	
	Original	Stemmed	Original	Stemmed
H	NA	NA	NA	NA
HM	0.931	0.927	0.920	**0.940**
HMR	0.933	0.923	0.904	0.919
HM1	0.931	0.927	0.920	**0.940**
HMR1	0.933	0.923	0.904	0.919
HMR2	0.933	0.923	0.904	0.919

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

	multi-ref		single-ref	
	Original	Stemmed	Original	Stemmed
H	0.784	0.776	0.785	0.782
HM	0.959	0.949	0.917	0.918
HMR	0.882	0.864	0.753	0.718
HM1	0.859	0.847	0.853	0.849
HMR1	**0.961**	0.952	0.921	0.914
HMR2	0.860	0.848	0.855	0.847

DUC-2003 ROUGE vs. Human Scores Pearson's Correlation

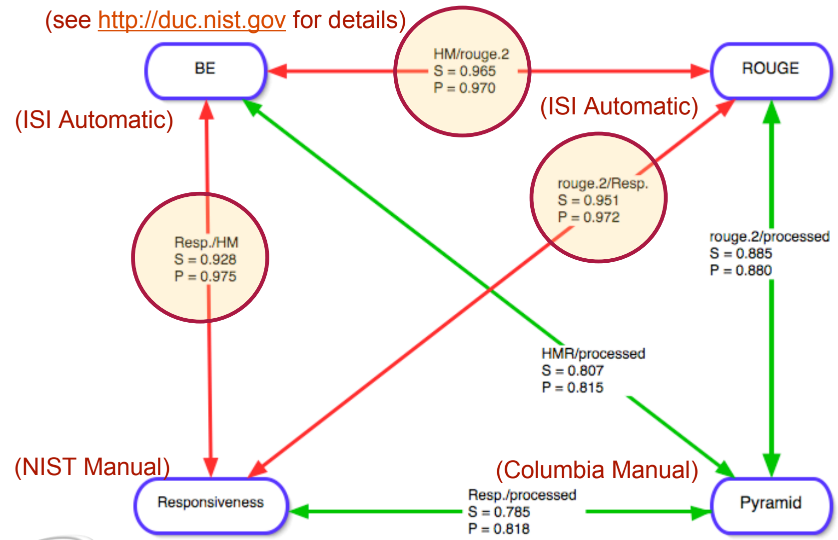
	multi-ref			single-ref		
	Original	Stemmed	Stopped	Original	Stemmed	Stopped
R1	0.619	0.609	0.773	0.622	0.611	0.786
R2	0.875	0.883	**0.921**	0.803	0.796	0.895
R3	0.872	0.869	0.880	0.684	0.669	0.687
R4	0.736	0.733	0.647	0.488	0.488	0.501
RL	0.547	0.533	0.726	0.539	0.508	0.729
RS4	0.811	0.817	0.886	0.744	0.754	0.885
RSU4	0.747	0.748	0.845	0.723	0.726	0.864

Correlation Analysis: DUC 2005

USC Viterbi School of Engineering

(S: Spearman's correlation; P: Pearson's correlation)

(see http://duc.nist.gov for details)



Conclusions

USC Viterbi School of Engineering

- BE-F consistently achieves over 90% Peason'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

USC Viterbi School of Engineering

- 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.

ISI Information Sciences Institute