BAE SYSTEMS

Statistical Evaluation of Information Distillation Systems

J.V. White, D. Hunter, J.D. Goldstein

BAE Systems, AIT (Advanced Information Technologies)

LREC 2008, Marrakech, Morocco
May 30, 2008

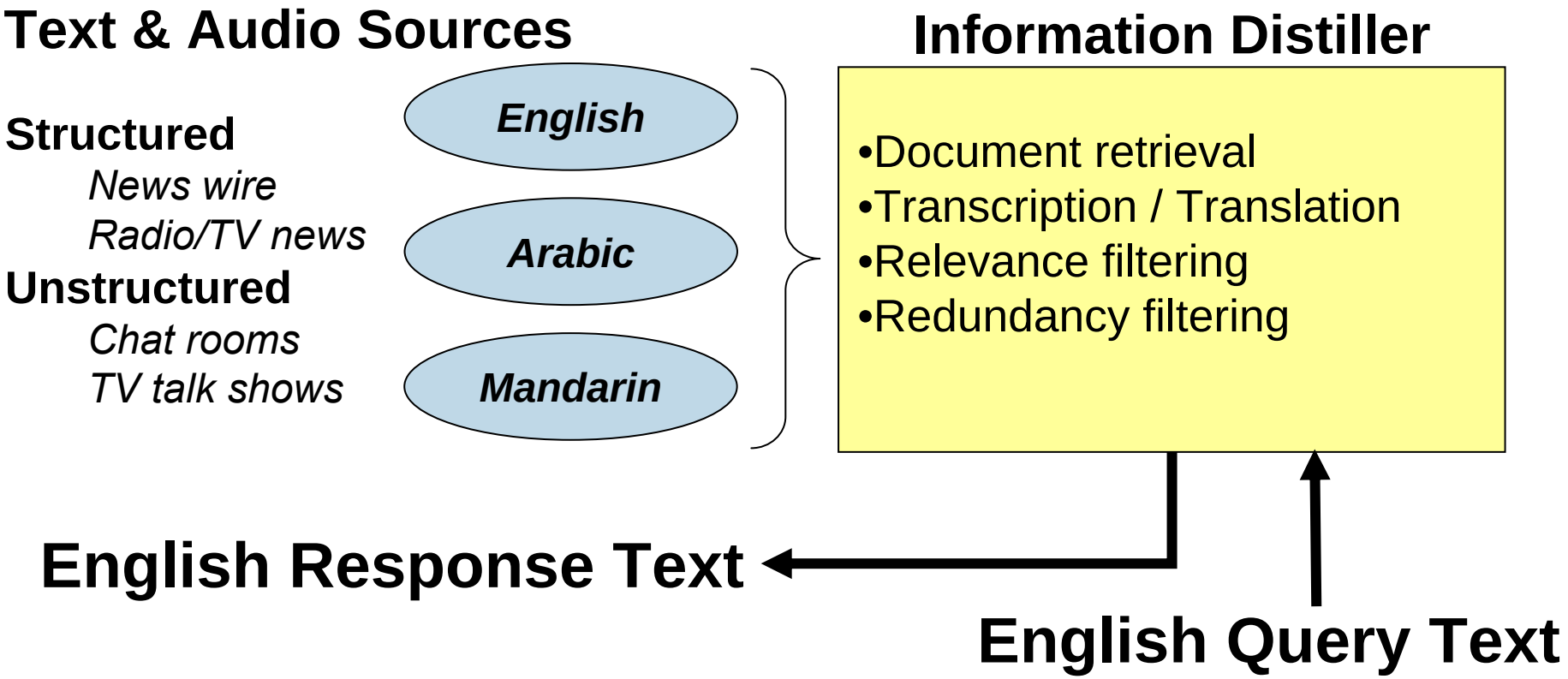
BAE SYSTEMS

Overview

- Information distillation
- Evaluation objectives
- Data analysis and statistical methodology
- Simple examples

BAE SYSTEMS

Information distillation



- English queries produce English response texts, even with foreign text/audio sources

BAE SYSTEMS

Acknowledgement

- Our evaluation methodology was developed for the GALE program (Global Autonomous Language Exploitation) under DARPA/IPTO support

- Our methodology may be used for other evaluations that share similar objectives
  - Evaluate unstructured response texts for information content
  - Penalize responses for irrelevance, redundancy, and missing information
  - Provide comprehensive statistical performance metrics (recall, precision, F-value, proficiency, …)

BAE SYSTEMS

System evaluation scope

- Our methodology focuses on overall *system* performance – not component performance
  - We evaluate the distiller response for relevance to the input query
  - The performance evaluation penalizes redundant, missing, and irrelevant information, as well as any gibberish in the response
  - We don't use transcription and translation metrics, but such errors reduce the system-level performance that we do measure
- This presentation doesn't describe our methodology for evaluating document citations
- Nor does this presentation address usability, readability, or utility metrics

BAE SYSTEMS

Technical approach

- Annotators divide the distiller's relevant response text into *information nuggets* and group these into *nugs*
  - *Nugs* are fuzzy sets of more-or-less equivalent nuggets
  - The *nuggets* are manually identified by annotators in GALE
  - In principle, nuggets may be automatically parsed
- Annotators analyze the information content of the nuggets for relevance to the query
- Annotation tasks include
  - Grouping nuggets into nugs based on their meanings
  - Assigning *Relevance weights* to the nugs
  - Assigning *Degrees of membership* to relatively imprecise nuggets that overlap more specific nuggets in meaning

BAE SYSTEMS

Nug 1

- Query: How are Joan and Bill related to each other?

Nug text	Nug relevance	Distiller ID	Nugget text	Degree of membership
They authored the book, *Evaluation Made Simple.*	1.0	A	They are joint authors.	0.5
		B	They authored the book, *Evaluation Made Simple.*	1.0
		C	*(No nugget provided.)*	0.0
		D	*(No nugget provided.)*	0.0

- The Nug is a fuzzy equivalence class of nuggets
- Meaning of Nug = meaning of its most precise Nugget

BAE SYSTEMS

Nug 2

- Query: How are Joan and Bill related to each other?

Nug meaning	Nug relevance	Distiller ID	Nugget text	Degree of membership
They wrote the paper, “Further thoughts on evaluation.”	1.0	A	They are joint authors.	0.5
		B	They wrote the paper, “Further thoughts on evaluation.”	1.0
		C	They wrote the paper, “Further thoughts on evaluation.”	1.0
		C	*Redundant nugget:* They wrote the paper, “Further thoughts on evaluation.”	1.0
		D	*(No nugget provided.)*	0.0

BAE SYSTEMS

Nug 3

- Query: Where does Joan live? (looking for address)

Nug meaning	Nug relevance	Distiller ID	Nugget text	Degree of membership
Joan lives in Rome, Italy.	0.5 (no address)	A	Joan lives in Italy.	0.5
		B	Joan lives in Rome, Italy.	1.0
		C	Joan lives in Italy's capital.	1.0
		C	*Redundant, imprecise nugget:* Joan lives in Italy.	0.5
		D	*(No nugget provided.)*	0.0

BAE SYSTEMS

Nug analysis for a set of queries

- Count the nugs and count the nuggets from each distiller being evaluated
- Compute statistics for each distiller
  - # *relevant* nugs
  - # *redundant* nuggets (more than one nugget in a nug)
  - # *missed* nugs that were found by other distillers
- Nug analysis determines whether or not different distillers provide nuggets that mean essentially the same thing, no matter how they are expressed

BAE SYSTEMS

Classical statistical methodology

- For each distiller, use a 2x2 contingency table and two indicator variables (*x,y)* to define four contingencies involving nugs and nuggets
  - *(x=1, y=1)* ***Relevant nug*** and **distiller contributes** a nugget to it
  - *(x=1, y=0)* ***Relevant nug*** and **distiller does not contribute** a nugget
  - *(x=0, y=1)* ***Irrelevant nug*** and **distiller contributes** a nugget to it
  - *(x=0, y=0)* ***Irrelevant nug*** and **distiller does not contribute** a nugget to it

BAE SYSTEMS

Classical contingency table for a distiller

	y = 0	*y = 1*
x = 0	# Other	# Wrong
x = 1	# Missing	# Right

This classical approach ignores the *relevance weight* of each nug, the *degree of membership* for each nugget, and fails to count *redundant* nuggets as being wrong

BAE SYSTEMS

Relevance and degree of membership

- To measure *relevance, redundancy,* and *degrees of membership*, define three fuzzy descriptors for each distiller
  - R_k = relevance weight of nug *k*
    - *R generalizes the relevance indicator x*
  - D_k = largest degree of membership in nug *k* (for distiller of interest)
    - *D generalizes the existence indicator y*
  - D_{kj} = degree of membership of the *j*-th redundant nugget in nug *k* (redundant nuggets have degrees of membership that do not exceed D_k)
  - All of these fuzzy measures take values on the unit interval [0, 1]

BAE SYSTEMS

Contingency table based on fuzzy counts

- The counts in the contingency table for a specified distiller satisfy these sums over all nugs

$$\#\ \text{Right} = \sum_k R_k D_k$$

$$\#\ \text{Wrong} = \sum_k \left((1 - R_k) D_k + \sum_j D_{kj} \right) + \text{Estimated \# Wrong in un-nuggetized text}$$

$$\#\ \text{Missing} = \sum_k R_k (1 - D_k)$$

$$\#\ \text{Other} = \sum_k (1 - R_k)(1 - D_k) + \text{Estimate of \# Other in the corpora}$$

  - Each nug contributes at most one count to the table
  - As $R \rightarrow x$ and $D \rightarrow y$, counting statistics $\rightarrow$ classical values
  - Fractional counts are distributed so that relevance and degree-of-membership contributions are statistically independent

BAE SYSTEMS

Estimating # wrong nugs in irrelevant text

- Irrelevant text is not nuggetized
- Therefore, the # Wrong (irrelevant) nugs is estimated from the total number of non-blank characters (# Char) in the distiller's response text

$$\#\ \text{Wrong} = \max\left(0, \frac{\#\ \text{Char}}{40} - \#\ \text{Right}\right)$$

- 40 is the average number of non-blank characters per nug (empirically determined)

BAE SYSTEMS

Classical performance metrics

	y = 0	*y = 1*
x = 0	# Other	# Wrong
x = 1	# Missing	# Right

	y = 0	*y = 1*
x = 0	$P_{xy}(0,0)$	$P_{xy}(0,1)$
x = 1	$P_{xy}(1,0)$	$P_{xy}(1,1)$

- Compute the *joint probability distribution* P_{xy} by normalizing the contingency table
- Classical performance metrics are functions of P_{xy}
  - Recall = $P_{y|x}(1,1) = P_{xy}(1,1) / P_x(1)$
  - Precision = $P_{x|y}(1,1) = P_{xy}(1,1) / P_y(1)$
  - F-value = 2 x Recall x Precision / (Recall + Precision)

BAE SYSTEMS

Proficiency metric

- *Proficiency* measures the fraction of information delivered by the distiller, relative to the total relevant information from all the distillers
- *Proficiency* = normalized mutual information between *x* and *y* indicator variables

$$\text{Proficiency} = \frac{I_{XY}}{H_X}$$

$$I_{XY} = \text{ mutual information} = \sum_x \sum_y P(x,y)\ \text{Log}_2 \frac{P(x,y)}{P(x)P(y)}$$

$$H_X = \text{ entropy} = -\sum_x P(x)\ \text{Log}_2 P(x)$$

BAE SYSTEMS

Interpretation of Proficiency

- Takes values on the unit interval [0, 1]
- Value 0 means that the distiller provides no relevant information
- Value 0.75 means that the distiller provides 75% of the relevant information delivered by all the distillers being considered

BAE SYSTEMS

Bayesian analysis

- For small samples, we recommend a Bayesian correction to the contingency table (CT)
- For example, add 1/4 pseudo count to each cell of the CT
  - This avoids zero counts and zero probabilities caused by small sample sizes
  - The performance metrics are then always well defined and take reasonable values, even with small amounts of evidence
  - With no data, the prior values for recall and precision are each ½, and the proficiency is 0

BAE SYSTEMS

Example performance metrics (1/2)

- We use contingency tables based on Nugs 1 – 3
- We assume the corpora contain on order of 10^5 nugs
- # Wrong in *irrelevant text* (not shown) is estimated from character counts

Estimated # Wrong

Distiller	# Wrong
A	1.50
B	1.00
C	0.75
D	0.00

BAE SYSTEMS

Example performance metrics (2/2)

Based on raw empirical probabilities

Distiller	Precision	Recall	Proficiency
A	0.417	0.500	0.400
B	0.625	1.000	0.909
C	0.154	0.333	0.234
D	(undefined)	0.000	0.000

These extreme values reflect the small sample size

Based on Bayesian probabilities

Distiller	Precision	Recall	Proficiency
A	0.429	0.500	0.400
B	0.611	0.917	0.811
C	0.200	0.375	0.271
D	0.500	0.083	0.066

The *Proficiency* provides rank ordering of distillers based on relative information content

BAE SYSTEMS

Conclusion

- Information distillers generate English response text from both structured and unstructured multilingual sources
- We have developed a statistical methodology for evaluating such distillers, which measures
  - Relevance
  - Redundancy
  - Recall
  - Precision
  - Proficiency (quantity of information provided)