Statistical Evaluation of Information Distillation Systems

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

BAE Systems, Advanced Information Technologies
6 New England Executive Park, Burlington, MA 01803
James.V.White@BAEsystems.com, Daniel.Hunter@BAEsystems.com, Jacob.Goldstein@BAEsystems.com

Abstract

We describe a methodology for evaluating the statistical performance of information distillation systems and apply it to a simple illustrative example. (An information distiller provides written English responses to English queries based on automated searches/transcriptions/translations of English and foreign-language sources. The sources include written documents and sound tracks.) The evaluation methodology extracts information *nuggets* from the distiller response texts and gathers them into fuzzy equivalence classes called *nugs*. The methodology supports the usual performance metrics, such as *recall* and *precision*, as well as a new information-theoretic metric called *proficiency*, which measures how much information a distiller provides relative to all of the information provided by a collection of distillers working on a common query and corpora. Unlike previous evaluation techniques, the methodology evaluates the relevance, granularity, and redundancy of information nuggets explicitly.

1. Information Distillers

An autonomous information distiller takes written queries as inputs, and in response, automatically gathers, transcribes, translates (if necessary), and distills relevant information from multilingual text and speech sources. The distiller outputs the distilled information in a readable document written in the same language as the query. The distiller also identifies all of the source files that support each fact or assertion in the distilled information. *Precise* distillers produce concise clean output: they avoid presenting *redundant*, *mistranscribed*, *mistranslated*, or *irrelevant* information. *Completely thorough* distillers miss nothing: they report all of the relevant information in the corpora being queried.

This paper discusses a methodology for statistically evaluating the *information content of distiller responses*. The handling of document citations, the usability, readability, and utility of the responses, and translation quality metrics are not discussed in this paper.

Although our evaluation methodology was developed to support the GALE (Global Autonomous Language Exploitation) program,[1] [2] it is also applicable to other evaluations that share similar objectives. The methodology is based on analyzing the nuggets of information contained in the distiller's response. The nuggets may either be manually produced by annotators, as they are in the GALE program or as in the original Pyramid approach for evaluating summaries (Nenkova-Passonneau, 2003), or the nuggets may be automatically extracted, as in (Zhou-Hovy, 2007; Zhou-Hovy, 2006). However, even if nuggets are extracted automatically, the statistical methodology described here does require some manual annotation because annotators must assign *relevance weights* to the relatively precise and specific nuggets, and they must assign *degrees of membership* to any relatively imprecise nuggets that partially overlap more specific nuggets.

2. Evaluation Objectives

A primary objective of GALE is to compare the performance of several automatic machine distillers with multilingual human distillers (who use only consumer software tools for word processing, reviewing audio, and file searching). Our evaluation methodology supports all of the standard statistical metrics for information extraction as well as several new ones, such as a *citation weighted F* metric, a *rightness* metric, and an information-theoretic *proficiency* metric, which are defined in Section 6.

A distiller is allowed to produce any readable response text that is consistent with the sources; there are no significant structural constraints on the distiller output. A valid response may be a sequence of direct quotes from the sources, or it may include paraphrases or summaries of source material. The distiller is free to use any kind of wording in its responses, as long as the resulting response is readable and free of redundancy. Therefore, a major objective of the evaluation is to evaluate the information content of *unstructured* responses. If two distillers use completely different wording but provide exactly the same information with the same redundancies, then their responses should be evaluated as being equivalent. In practice, one response may be more readable than the other, but this issue is ignored in the present methodology.

Our evaluation for GALE combines scoring of both query responses and document retrieval, but this paper deals only with the analysis of the query responses and does not address the evaluation of document citations.

There have been a number of formal evaluations of document retrieval systems (TREC, 2005), and while these evaluations have some points of contact with the evaluation described in this paper, the more relevant comparison is with evaluations of query answering systems such as those done in the query answering track in TREC-2005

[1] This material is based upon work supported by the Defense Advanced Research Projects Agency DARPA/IPTO, Global Autonomous Language Exploitation, ARPA Order No. V018, Program Code No. 5M30, issued by DARPA/CMO under Contract #HR0011-06-C-003. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency or the U.S. Government.

[2] http://www.arpa.mil/ipto/programs/gale.

(Boorhees-Dang, 2005). Our evaluation methodology has some features that distinguish it from previous evaluations of query answering systems. A major difference is the open-ended nature of the queries. Distillers in Gale were required to answer queries such as "Provide information about [EVENT]" or "Give a biography of [X]", where a wide variety of different types of information might be relevant. In contrast, almost all the queries in TREC-2005 had answers that were either a single item or a list of items. (For example, "When was Hong Kong returned to Chinese sovereignty" or "Which other countries formally congratulated China on the return?")

The open-ended nature of the queries in Gale has implications for the method of evaluation. The relevance of a response becomes harder to judge, and it becomes more important to have a finer grained measure of relevance than that used for more narrowly focused queries. For open-ended queries, the specificity of a response also becomes important, and specificity is a different dimension than relevance. ("John was in Italy" is less specific than "John was in Rome, Italy," but both are fully relevant to the question "Where has John been?") In the TREC-2005 query answer track, qualitative notions of specificity and relevance were combined in the evaluation metrics in such a way that their individual contributions to response quality could not be determined. One of the primary goals in the development of our evaluation methodology was to find a principled way of independently measuring relevance and specificity and combining the two in calculations of precision, recall, and other evaluation metrics.

3. Experimental Design

The GALE Phase-2 evaluation corpora[3] include source files from three languages (English, Farsi, and Mandarin), two media types (text and audio), unstructured formats (Internet chat rooms and TV talk shows), and structured formats (news wire and radio/TV news shows). Therefore the data space is partitioned into twelve cells, each corresponding to a different (language, media type, structure class) trio. For each of these cells, we generate four or five queries by putting the names of persons, organizations, or countries into the blank fields of query templates.

4. Data Analysis

Each distiller produces *snippets* of text in their responses. A snippet is defined to be a sentence or meaningful sentence fragment that answers, or partially answers, the query. We analyze each snippet to determine how many *nuggets* of information it contains. Nuggetization details for GALE are discussed in (Babko-Malaya, 2008) and will not be discussed here. A single distiller, or a pair of distillers, may provide two nuggets that mean nearly the same thing, although their wordings are different. These two nuggets are grouped together in a set, which we call a *nug*.[4] If another distiller provides a third nugget that means essentially the same thing as the first two, then we put it into the nug too. In this way, we can determine

- how many nuggets of information each distiller provides
- whether or not a distiller provides redundant nuggets (more than one nugget in a nug)
- whether or not a distiller has missed nuggets that were found by other distillers
- whether or not two different distillers provide nuggets that mean essentially the same thing.

The *meaning of a nug* is defined to be the meaning of its most precise nugget. Some nugs are better than others because they are more relevant to the query. To measure pertinence to the query, annotators assign a *relevance weight* to each nug. Moreover, some of the nuggets inside a nug may be better than other nuggets in it because their meanings more nearly match the nug's meaning. To reflect these granularity differences, annotators assign a *degree of membership* to each nugget in a nug. Degrees of membership and relevance weights are numbers in the unit interval [0, 1]. If there is only one nugget in a nug, its degree of membership is exactly 1, because this nugget conveys the exact meaning of the nug. Annotators assign every nugget to a nug. Infrequently they assign a nugget to more than one nug, and then only if its meaning overlaps more than one nug. A nugget in more than one nug has degrees of membership in the nugs that sum to at most 1. This guarantees that each nugget contributes no more than one count total to the contingency table.

In summary, the relevance weight of a nug measures how relevant the meaning of the nug is to answering the query. The degree of membership assigned to a nugget measures how nearly the meaning of the nugget matches the meaning of its nug. Those nuggets that are less precise than others in the nug have degrees of membership less than 1.

5. Statistical Methodology

Our statistical methodology is based on contingency tables, which follow standard practice. However, to capture information provided by relevance weights and degrees of membership, we use fuzzy set theory to compute the counts in the contingency table.

To measure the information content of the responses from a single distiller, say distiller A, we form a contingency table in which we count (or estimate) the following statistics.

 # Right Nuggets A nugget is *right* if it is relevant to answering the query and is not redundant with another nugget from distiller A.

 # Wrong Nuggets A nugget is *wrong* if it is not relevant to answering the query or if it is relevant but redundant. Since annotators nuggetize only relevant snippets, we *estimate* the number of irrelevant nuggets based on character counts. Redundant nuggets, on the other hand, are directly counted.

 # Missing Nuggets A nugget is *missing* whenever distiller A fails to contribute a nugget to a relevant nug that is

[3] The corpora are provided by the Linguistic Data Consortium. http://projects.ldc.upenn.edu/gale

[4] In terms of fuzzy set theory, a nug is a fuzzy equivalence class (Dumitrescu et al, 2000).

supported by nuggets from at least one other distiller. Missing nuggets, by definition, have a degree of membership equal to zero.

 # Other Nuggets The *other* nuggets are all those unobserved nuggets in the corpora that are neither contained in any distiller's snippets nor relevant to the query. Most performance metrics don't depend on this count. However, the information-theoretic metric *proficiency* does. (Proficiency is defined by Eq.(12).) An order-of-magnitude estimate of this count (based on character counts) is sufficient for computing proficiency in large corpora.

The contingency table for distiller A is the matrix of nugget counts, which records how many nuggets are

- Relevant ($x = 1$) and included in the distiller's response ($y = 1$)
- Irrelevant or redundant ($x = 0$) and included in the distiller's response ($y = 1$)
- Relevant ($x = 1$) and excluded from the distiller's response ($y = 0$)
- Irrelevant ($x = 0$) and excluded from the distiller's response ($y = 0$).

The contingency table has the following layout

$$\begin{array}{c} \\ x=0 \\ x=1 \end{array} \begin{array}{c} \begin{array}{cc} y=0 & y=1 \end{array} \\ \begin{bmatrix} \#\,\text{Other} & \#\,\text{Wrong} \\ \#\,\text{Missing} & \#\,\text{Right} \end{bmatrix} \end{array}.$$

We construct this table for *each distiller-query pair*. Suppose that we are dealing with distiller A. We use the following formulas, in which summations extend over all of the nugs produced by all human and machine distillers being evaluated, R_k is the relevance weight for nug k, D_k is the degree of membership of the most precise nugget contributed to nug k by distiller A (note that $D_k = 0$ if A has no nugget in the nug):

$$\#\,\text{Right}_A = \sum_k R_k D_k, \tag{1}$$

$$\#\,\text{Wrong}_A = \sum_k (1 - R_k) D_k + \#\,\text{Redundant}_A(k) + \text{Estimated \# Wrong in un-nuggetized text}, \tag{2}$$

$$\#\,\text{Missing}_A = \sum_k R_k (1 - D_k), \tag{3}$$

$$\#\,\text{Other}_A = \sum_k (1 - R_k)(1 - D_k) + \text{Estimate based on character counts}. \tag{4}$$

Here

$$\#\text{Redundant}_A(k) = \sum_j D_{kj}$$

is the effective number of redundant nuggets from distiller A in nug k, and D_{kj} is the degree of membership of distiller A's jth redundant nugget in the kth nug. These formulas guarantee that each nug contributes no more than one count per nug to the contingency table. Moreover, when a count is split among two or more cells, the split is defined so that the redundancy and degree-of-membership contributions to it are statistically independent of each other.

6. Performance Metrics

6.1. Empirical Probability Model

We compute statistical performance metrics for distiller A in two steps. First the contingency table is used to specify a joint probability matrix P_{XY} for the two indicator variables x and y,

$$P_{XY} = \begin{bmatrix} \Pr(x=0, y=0) & \Pr(x=0, y=1) \\ \Pr(x=1, y=0) & \Pr(x=1, y=1) \end{bmatrix}. \tag{5}$$

The empirical joint probability model $P(x, y)$ is computed by dividing each count in the contingency table by the total number of counts. From this model, many performance metrics may be computed. For example,

$$\text{precision} = P(x = 1 \mid y = 1), \tag{6}$$

$$\text{recall} = P(y = 1 \mid x = 1), \tag{7}$$

$$\text{rightness} = P(y = x = 1)/(1 - P(y = x = 0)), \tag{8}$$

$$\text{accuracy} = P(y = x = 0) + P(y = x = 1), \tag{9}$$

$$\text{CW } F\text{-value} = \frac{2 \times \text{precision} \times \text{CW recall}}{\text{precision} + \text{CW recall}}. \tag{10}$$

Comments on *CW F-value* This metric penalizes the distiller if its document citations are deficient in either citation-recall or citation-precision. The *citation-weighted CW recall* in the definition of CW F-value is not expressible in terms of the joint probability matrix $P(x, y)$, because CW recall is a weighted sum of the mean degrees of membership of the distiller's nuggets,

$$\text{CW recall} = \frac{1}{N} \sum_{n=1}^{N} \bar{D}_n \sqrt{F_n^C}, \tag{11}$$

where N is the total number of nugs, $\bar{D}_n$ is the mean degree of membership of the distiller's nuggets in nug n, and F_n^C is the F-value for the document citations provided by the distiller for the nuggets in nug n. We use the square root of F_n^C to soften the impact of low F-values on citation-weighted recall.

Comments on *rightness* *Rightness* is the fraction of the observed nuggets (right, wrong, and missing) that are expected to be *right*. The *rightness* metric is a lower bound on both precision and recall. In fact, *rightness* equals precision if recall equals 1, and *rightness* equals recall if precision equals 1. In terms of sample counts, *rightness* = # right / (# right + # wrong + # missing).

Comment on *accuracy* This is an uninteresting metric for large corpora, because $P(y = x = 0)$ is very close to 1, and hence, *accuracy* is very close to 1 no matter what the distiller does.

***Proficiency* metric** We also use a calibrated and normalized information-theoretic metric, which we call *proficiency* (White et al, 2004),

$$\text{proficiency} = \begin{cases} I_{XY}/H_X & \text{if } H_X > 0, \\ 0 & \text{if } H_X = 0 \text{ and } H_Y > 0, \\ 1 & \text{if } H_X = 0 \text{ and } H_Y = 0. \end{cases} \quad (12)$$

where I_{XY} is the mutual information between x and y, while H_X and H_Y are the entropies of x and y. These are defined as

$$I_{XY} = \sum_x \sum_y P(x, y) \log_2 \frac{P(x, y)}{P(x)P(y)}, \quad (13)$$

$$H_X = -\sum_x P(x) \log_2 P(x) \qquad \text{(and the same for } Y\text{)}, \quad (14)$$

in which we set "$0 \log_2 0$" equal to 0.

Proficiency takes values on the unit interval $[0, 1]$. Like the F-value, *proficiency* is a metric that penalizes a distiller for both *missing* and *wrong* information nuggets. However, unlike the F-value, *proficiency* also takes into account how large the test corpora are. The *proficiency* measures the fraction of information that is actually delivered by the distiller, relative to the total information delivered by all of the distillers. Thus, if a distiller A delivers 75% of the information provided by all of the distillers in the evaluation, then distiller A has a *proficiency* of 0.75.

6.2. Bayesian Probability Model

A practical problem arises when the contingency table contains one, or more, zeros because certain possible outcomes were not observed in the experiment. Such zeros are typically caused by small sample sizes. If the zeros appear in the probability model, then some metrics may be undefined. To avoid this impasse, we recommend using a Bayesian approach to specifying the probability matrix $P(x, y)$, which prevents zeros from appearing in the final posterior model. The Bayesian approach is particularly suited to comparing machine distillers with human distillers, because the effects of the prior densities on ratios of machine-to-human metrics is quite reasonable.

In the Bayesian approach,[5] before any experimental results are entered, all distillers are assigned the same prior non-informative contingency table:

$$C_{\text{prior}} = \begin{bmatrix} \#\,\text{Other} & 1 \\ 1 & 1 \end{bmatrix},$$

which initially puts all of the distillers on the same neutral playing field.

The counts computed from formulas (1) thru (3) are then added to their corresponding prior counts. The resulting contingency table, when normalized to produce a joint probability matrix, yields a Bayesian posterior model,[6] which contains strictly non-zero probabilities.

7. Distillation Examples

The GALE program is in-progress, and DARPA has not released any official results yet. Therefore, we provide examples of applying our statistical methodology to some hypothetical distillation data, which we crafted to support our pedagogical objectives. Because of space constraints, we consider just four hypothetical distillers, A, B, C, and D, although the methodology handles any number.

7.1. Example Query

Consider the query "How are Joan and Bill related to each other?" Table 1 contains examples of information nuggets from four responding distillers. Each nug has the indicated *relevance* to the query, and each nugget has a *degree-of-membership* (DM) in each related nug. Distiller A's nugget "They are joint authors." is imprecise and overlaps both nugs in meaning. Therefore, the annotator assigned this nugget a DM = 0.5 in both nugs. Distiller B provided two nuggets that precisely fit into their respective nugs. Therefore, the DMs are both equal to 1. Distiller C failed to provide a nugget that fits into the first nug, so its DM for this nug is zero. Distiller C also made a redundancy error: C reported two identical nuggets for the second nug. Distiller D failed to report any nuggets that fit the two nugs and so received DMs of zero for each missing nugget.

7.2. Another Example Query

Consider the query, "Where is Joan?" In this example, we are assuming that the street address of her location is being requested. Table 2 presents nugs and nuggets for this query, based on the responses from the four distillers. All the responses are inadequate to get full credit. The best nuggets come from distillers B and C, which identify the city and country of her location. So they both have DM = 1.0 for these best nuggets. However, distiller C makes the error of providing a redundant, imprecise nugget, which has DM = 0.5. The best nuggets fail to provide the desired street address of Joan's location, so the annotator assigns a relevance of only 0.5 to the resulting nug.

7.3. Example Contingency Tables

The contingency table for a specific distiller is formed by (1) computing the incremental contributions from each nug according to the relevance and DM metrics in Table 2 and (2) then adding the incremental contributions together to produce the table. Table 3 shows the incremental contributions to the contingency tables for each distiller. These increments (fractional counts) are computed by using equations (1) – (4). Note that the incremental contributions for each nugget sum to exactly 1 count, except for redundant nuggets, which contribute only to the #wrong count according to their degree of membership in the nug.

[5] A standard Bayesian setup puts non-informative prior Dirichlet probability densities on the three entries in the lower right corner of $P(x, y)$, and a number in the upper left corner, which is an order-of-magnitude estimate of the number of "other" nuggets in the evaluation corpora.

[6] In the Bayesian analysis, the final probabilities in $P(x, y)$ are expected values from the posterior Dirichlet density.

Table 1: Nuggets and nugs for example query.

Query: How are Joan and Bill related to each other?				
Nug	**Relevance**	**Distiller**	**Nugget**	**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
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

Table 2: Nuggets and nugs for another example query.

Query: Where is Joan?				
Nug	**Relevance**	**Distiller**	**Nugget**	**Degree of Membership**
Joan is in Rome, Italy.	0.5	*A*	Joan is in Italy	0.5
		B	Joan is in Rome, Italy	1.0
		C	Joan is in Italy's capital city.	1.0
		C	*Redundant, imprecise nugget:* Joan is in Italy.	0.5
		D	*(No nugget provided)*	0.0

By adding the incremental counts from Table 3, we obtain the following *partial* contingency tables.

$$\text{Distiller}A: \begin{bmatrix} 0.25 & .25 \\ 1.25 & 1.25 \end{bmatrix} \quad (15)$$

$$\text{Distiller}B: \begin{bmatrix} 0.00 & 0.50 \\ 0.00 & 2.50 \end{bmatrix} \quad (16)$$

$$\text{Distiller}C: \begin{bmatrix} 3.00 & 2.00 \\ 1.00 & 0.00 \end{bmatrix} \quad (17)$$

$$\text{Distiller}D: \begin{bmatrix} 0.50 & 0.00 \\ 2.50 & 0.00 \end{bmatrix}. \quad (18)$$

The final *full* contingency tables are obtained from the partial ones by making two additions to each table: (1) adding an order-of-magnitude estimate of the number of *Other* nugs in the test corpora to the #Other counts; and (2) adding an estimate W_e of the #Wrong counts contributed by the irrelevant text that was submitted by each distiller but not nuggetized. For the present example, we assume that there are a total of roughly 10^5 nugs in the corpora. The estimated number of nugs in irrelevant text is computed as

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

where #char is the number of characters in the irrelevant text, and the mean density of nuggets in the irrelevant text from each distiller is assumed to be 40 non-blank characters/nug. In practice, this density is determined by nuggetizing a body of irrelevant text for the distillers and determining the density empirically. Suppose that the distillers submitted the amount of irrelevant text shown in Table 4. The resulting full contingency tables are

Table 4: Estimating #Wrong nugs in irrelevant text.

Distiller	**#Characters**	**Estimated #Wrong**
A	60	1.50
B	40	1.00
C	30	0.75
D	0	0.00

$$\text{Distiller}A: \begin{bmatrix} 10^5 & 1.75 \\ 1.25 & 1.25 \end{bmatrix} \quad (19)$$

$$\text{Distiller}B: \begin{bmatrix} 10^5 & 1.50 \\ 0.00 & 2.50 \end{bmatrix} \quad (20)$$

$$\text{Distiller}C: \begin{bmatrix} 10^5 & 2.75 \\ 1.00 & 0.00 \end{bmatrix} \quad (21)$$

$$\text{Distiller}D: \begin{bmatrix} 10^5 & 0.00 \\ 2.50 & 0.00 \end{bmatrix}. \quad (22)$$

The empirical joint probability distribution P_{XY} for each distiller is computed by normalizing each full contingency matrix.

7.4. Example Performance Metrics

Given the raw empirical probability distribution P_{XY}, the precision, recall, rightness, and proficiency metrics may be computed for each distiller by using equations (6)-(8) and (12) – (14. The metrics are displayed in Table 5. The *precision* for distiller *D* is NaN (not a number) because it is undefined: there is no empirical evidence whatsoever regarding

Table 3: Incremental Contributions to Contingency Tables.

Query: How are Joan and Bill related to each other?
Nug: They authored the book, *Evaluation Made Simple.*
Relevance: 1.0

Distiller	Nugget	Degree of Membership	#Right	#Wrong	#Missing	#Other	#All
A	They are joint authors.	0.5	0.50	0.00	0.50	0.00	1.00
B	They authored the book, *Evaluation Made Simple*.	1.0	1.00	0.00	0.00	0.00	1.00
C	*(No nugget provided)*	0.0	0.00	0.00	1.00	0.00	1.00
D	*(No nugget provided)*	0.0	0.00	0.00	1.00	0.00	1.00

Query: How are Joan and Bill related to each other?
Nug: They wrote the paper, "Further thoughts on evaluation."
Relevance: 1.0

Distiller	Nugget	Degree of Membership	#Right	#Wrong	#Missing	#Other	#All
A	They are joint authors.	0.5	0.50	0.00	0.50	0.00	1.00
B	They wrote the paper, "Further thoughts on evaluation."	1.0	1.00	0.00	0.00	0.00	1.00
C	They wrote the paper, "Further thoughts on evaluation."	1.0	1.00	0.00	0.00	0.00	1.00
C	*Redundant precise nugget:* They wrote the paper, "Further thoughts on evaluation."	1.0	0.00	1.00	0.00	0.00	1.00
D	*(No nugget provided)*	0.0	0.00	0.00	1.00	0.00	1.00

Nug: Joan is in Rome, Italy.
Relevance: = 0.5 (because her street address is needed)

Distiller	Nugget	Degree of Membership	#Right	#Wrong	#Missing	#Other	#All
A	Joan is in Italy.	0.5	0.25	0.25	0.25	0.25	1.00
B	Joan is in Italy's capital city.	1.00	0.50	0.50	0.00	0.00	1.00
C	Joan is in Italy's capital city.	1.00	0.50	0.50	0.00	0.00	1.00
C	*Redundant, imprecise nugget:* Joan is in Italy.	0.5	0.0	0.50	0.00	0.00	0.50
D	*(No nugget provided)*	0.0	0.00	0.00	0.50	0.50	1.00

the precision of this distiller. Also, *all* of the metrics for Distiller *C* are zeros, which is an extreme conclusion based on very little evidence. These undesirable, extreme, and unrealistic values are caused by the use of empirical probabilities when the sample sizes are too small for a classical approach to make much sense. The Bayesian approach in contrast, wherein 1 count is added to each cell in each full contingency table prior to normalization, provides well-defined metrics that do not overstate what is being deduced from small data samples. The resulting more realistic metrics are shown in Table 6. All of the resulting metrics assume moderate values, and the *proficiency* values provide a clear rank ordering of the distillers. Such a rank ordering was not available in Table 5, because both distillers *C* and *D* had zero proficiencies. Note that the *rightness* provides a different rank ordering than the *proficiency*. This difference arises because *rightness* tends to put the most weight on whichever of *recall* and *precision* is the smaller. In contrast, *proficiency* tends to put more weight on *recall*, because recall is most directly connected with finding information, while *precision* is more directly connected with user satisfaction at seeing clean output from the distiller.

Table 5: Performance metrics based on raw empirical probabilities.

Distiller	Precision	Recall	Rightness	Proficiency
A	0.417	0.500	0.294	0.400
B	0.625	1.000	0.625	0.909
C	0.000	0.000	0.000	0.000
D	NaN	0.000	0.000	0.000

Table 6: Performance metrics based on Bayesian probabilities.

Distiller	Precision	Recall	Rightness	Proficiency
A	0.450	0.500	0.310	0.399
B	0.583	0.778	0.500	0.665
C	0.211	0.333	0.148	0.238
D	0.500	0.222	0.182	0.176

8. Conclusions

We have described a methodology for evaluating the statistical performance of information distillation systems. The methodology assumes that annotators have identified relevant and possibly redundant information nuggets in the

responses from the distillers. The methodology also assumes that the nuggets have been collected into equivalence classes called nugs, which contain nuggets conveying essentially the same information with possibly differing degrees of precision. The methodology provides a direct means for distinguishing between distillers based on the relevancy of the information they provide, the density of relevant information in their response text, and the quantity of undesirable redundant information. Moreover, the methodology gives proper credit to a distiller, regardless of the exact wording it uses to express relevant information. Some simple pedagogical examples show how we apply the methodology in practice.

9. Acknowledgements

The authors acknowledge support for this work under the GALE program, which is funded by IPTO, Defense Advanced Research Agency. The authors also acknowledge the substantial technical contributions made by the program manager, Dr. Joseph Olive.

10. References

O. Babko-Malaya. 2008. "Annotation of Nuggets and Relevance in GALE Distillation Evaluation." *Proceedings LREC 2008*. Marrakech, Morocco.

D. Dumitrescu, B. Lazzerini, and L. C. Jain. 2000. *Fuzzy Sets and their Application to Clustering and Training.* CRC Press, Boca Raton.

A. Nenkova and R. Passonneau. 2003. *Evaluating Content Selection in Summarization: The Pyramid Method.* Technical Report CUCS025-03. Columbia University. http://citeseer.ist.psu.edu/passonneau03evaluating.html.

TREC 2005. http://trec.nist.gov/.

E. M. Voorhees and H. T. Dang. 2005. "Overview of the TREC 2005 Question Answering Track." *Proceedings of the Fourteenth Text Retrieval Conference (TREC 2005)*.

J. V. White, S. Steingold, and C. G. Fournelle. 2004. "Performance Metrics for Group-Detection Algorithms." *Proceedings of Interface 2004.* Baltimore, MD, May.

L. Zhou and E. H. Hovy. 2006. "A Semi-automatic Evaluation Scheme." *Proceedings of the DARPA GALE PI workshop*. San Francisco, CA. http://www.isi.edu/natural-language/people/hovy/papers/07DARPA-ParaEval.pdf.

L. Zhou and E. H. Hovy. 2007. "A Semi-Automated Evaluation Scheme: Automated Nuggetization for Manual Annotation." *Proceedings of the Human Language Technology / North American Association of Computational Linguistics conference (HLT-NAACL 2007)*. Rochester, NY.