EVALUATION OF DOCUMENT CITATIONS IN PHASE 2 GALE DISTILLATION

Olga Babko-Malaya, Dan Hunter, Connie Fournelle, Jim White
BAE Systems
6 New England Executive Park
Burlington, MA 01803
E-mail: olga.babko-malaya/daniel.hunter/connie.fournelle/james.v.white@baesystems.com

Abstract

The focus of information retrieval evaluations, such as NIST's TREC evaluations (e.g. Voorhees 2003), is on evaluation of the information content of system responses. On the other hand, retrieval tasks usually involve two different dimensions: reporting relevant information and providing sources of information, including corroborating evidence and alternative documents. Under the DARPA Global Autonomous Language Exploitation (GALE) program, Distillation provides succinct, direct responses to the formatted queries using the outputs of automated transcription and translation technologies. These responses are evaluated in two dimensions: information content, which measures the amount of relevant and non-redundant information, and document support, which measures the number of alternative sources provided in support of reported information. The final metric in the overall GALE distillation evaluation combines the results of scoring of both query responses and document citations. In this paper, we describe our evaluation framework with emphasis on the scoring of document citations and an analysis of how systems perform at providing sources of information.

1. Introduction

This paper presents an approach to the evaluation of document citations in the Phase 2 DARPA Global Autonomous Language Exploitation (GALE) Program Distillation evaluation. The purpose of GALE Distillation evaluation is to quantify the amount of relevant and non-redundant information a distillation engine is able to produce and to compare that amount of information to the amount of information gathered by a bilingual human using commonly available state-of-the-art tools. GALE engines distill data in response to a formatted query from audio and text sources in English, Chinese, and Arabic; and they produce English-only responses using translations and transcriptions. GALE systems also report citations (i.e. sources of information), including alternative sources and corroborative evidence. As part of the evaluation, annotators judge the relevance and novelty of the returned responses, as well as check document citations in order to evaluate whether the content of the response accurately reflects information in the source documents. The final metric in the overall GALE distillation evaluation combines the results of scoring of both query responses and document citations. In this paper, we describe our evaluation approach with emphasis on annotation and scoring of document citations and an analysis of how systems perform at providing sources of information.

The citation checking task is done for English and foreign sources. As we discuss in the paper, the output of the citation checking task is used not only to compute the document support metrics, but also plays an important role for scoring of information content of query responses. During Phase 2 Distillation evaluation, in 30% of the cases, information in the snippet was not sufficient to make a decision on whether the snippet was relevant as an answer to the query, and annotators had to consult the original document to make this decision. The results of the evaluation show that, when compared to human performance, systems are able to analyze a much larger collection of documents than humans, and, therefore, systems benefit from including document citation metrics in the evaluation. Additionally, the citation checking task was also needed to identify errors in reported information due to incorrect machine translations, since the answer key pooled distiller responses. Without citation checking, errors could become part of the answer key, incorrectly penalizing other systems that did not make the same error.

The paper proceeds as follows. Section 2 presents an overview of the GALE Phase 2 Distillation evaluation. Section 3 discusses the citation checking requirements, and Section 4 presents annotation tasks. Sections 5 and 6 give an overview of scoring metrics, and Section 7 discusses the impact of citation checking on the final results of the evaluation.

2. GALE Phase 2 Distillation Overview

The queries in GALE Distillation conform to templates, which contain argument variables that range over events, topics, people, organizations, locations, and dates. In Phase 2 of the program, the set of seventeen templates included queries such as LIST FACTS ABOUT [event], FIND STATEMENTS MADE BY OR ATTRIBUTED TO [person] ON [topic(s)], DESCRIBE THE ACTIONS OF [person] DURING [date] TO [date]. Distillers produce English-only snippets in response to these queries. During evaluation, annotators create *nuggets*, or atomic pieces of information, out of relevant text, and map them to equivalence classes, called *nugs*.

For example, the following snippet is retuned in response to the query DESCRIBE ATTACKS IN [Kosovo]

***Snippet**: A UN policeman was killed late Sunday on the motorway between Leposavic and Mitrovica, some 55 kilometers north of the capital Pristina*

Annotators select three nuggets, which are indicated by double brackets below. The nuggets are created out of each core clause that includes the verb and its arguments, as well as temporal, locative, causative and other types of modifiers.

***Nugget 1**: [[A UN policeman was killed]] late Sunday on the motorway between Leposavic and Mitrovica, some 55 kilometers north of the capital Pristina*

***Nugget 2**: A UN policeman was killed [[late Sunday]] on the motorway between Leposavic and Mitrovica, some 55 kilometers north of the capital Pristina*

***Nugget 34**: A UN policeman was killed late Sunday [[on the motorway between Leposavic and Mitrovica, some 55 kilometers north of the capital Pristina]]*

It is certainly possible to break this snippet down into much smaller pieces, for example, by breaking down the locative expression above into more fine-grained nuggets (e.g. *"on the motorway", "between Leposavic and Mitrovica"*, and "*some 55 kilometers north of the capital Pristina").* However, in order to provide the level of granularity which corresponds to possible answers to the queries, as well as to simplify the annotation task, annotators were instructed to select the maximal extent of locative, temporal, and other types of modifiers. See (Babko-Malaya, 2008) for further discussion of the nuggetization rules in GALE Phase 2 Distillation evaluation.

Nugs are collections of nuggets from different distillers and from different source documents. Nuggets are put into the same nug if they are semantically equivalent or one nugget is more specific than the other. For example, the nugget "*Event E happened on [[August 27, 1991]]"* would be in the same nug as the nugget "*Event E occurred in [[August of 1991]]"* because one gives more precise information about the time of occurrence of the event than does the other. The semantic content of the nug itself can be taken to be the semantic content of the most precise nugget in the nug. See (White el al, 2009) for further discussion of the process of clustering nuggets into nugs.

Each nugget was annotated with a number between 0 and 1, called *nug degree of membership*, indicating its degree of specificity compared with the semantic content of the nug. In addition, the nugs themselves were annotated by their relevance to the query (their *degree of relevance*). These numbers were used to compute *information recall, precision*, and *F-value* for the response of a distiller to a query, as discussed in section 6 below. See also (White el al, 2008) for a more detailed discussion of our approach to computing information metrics.

We computed a parallel set of metrics for citations. The citation metrics measure the performance of the distillers in providing citation support for their query responses. To compute citation metrics, we evaluated the content of each nugget to verify that it accurately reflected information in the source document. *This specific task, called the 'citation checking task', is the focus of this paper.*

3. Citation Requirements

The goal of the citation checking task is to verify that snippets provided by the distiller, which included transcriptions and translations of foreign materials, are indeed supported by the citations listed in support of these snippets. Whereas human distillers provided only the document supporting their response, machine distillers were required to report the following:

- *Snippet Chunks*, which are excerpts of snippets (also in English text) that are supported by source citations. Chunks do not need to be self-sufficient (*i.e.*, interpretable independent of any other resource) and may contain non-contiguous text strings.
- *Citations*, which are the sources for the specific snippet chunks. Each citation indicates the chunk it supports, and includes the source document from which it originates.

Snippet chunks made it possible to provide more than one source for any snippet, where different sources may support different parts of the snippet. For example, the two citations below do not fully support the snippet individually, but each supports some of the information contained in the snippet:

***Snippet**: A UN policeman was killed late Sunday on the motorway between Leposavic and Mitrovica, some 55 kilometers north of the capital Pristina*

> ***Chunk 1**: A UN policeman was killed late Sunday on the motorway*
> ***Citation 1**: Menon became the first UN policeman to die in the line of duty in Kosovo when he was ambushed late Sunday on a motorway in northern Kosovo.*
>
> ***Chunk 2**: A UN policeman was killed late Sunday some 55 kilometers north of the capital Pristina*
> ***Citation 2**: Satish Menon, 43, from India's southern state of Kerala, was killed by sniper fire shortly before midnight Sunday while traveling in a U.N. police car near the village of Slatina, some 55 kilometers north of the capital, Pristina, police said*

By providing “snippet excerpts” or chunks, distillers indicated which part of the snippet is being supported by that citation.

4. Annotation

4.1 Chunk Degrees of Membership and Degree of Support

As part of the citation checking task, annotators verified that nuggets are indeed supported by the citations provided by the distiller. This involves two steps:

1. Judging each nugget’s *degree of chunk membership (cDM), and*
2. Judging the *degree of support* (DS) the citation provides to the content of the nugget contained in the chunk

The reason for breaking down the evaluation of nugget support into these two steps is that a nugget need not correspond exactly to a chunk provided by the distiller, as is illustrated by the examples below:

Nugget: *Menon was killed [[in northern Kosovo]]*
Chunk: *in … Kosovo (cDM=0.5)*
Citation*: the attack fell last night in Kosovo and that led to the death of one of United Nations policemen*
(DS=1)

Nugget: *[[A United Nations policeman has been shot dead]] in a sniper attack north of the capital Pristina*
Chunk: *… policeman has been shot dead … north of the capital Pristina (cDM=0.8)*
Citation*: The officer was killed between Leposavic and Mitrovica some 55 kilometers north of the capital Pristina (DS=1)*

In the first example, the nugget specifies a location. The citation fully supports the provided chunk but it does not fully support the nugget since the location in the nugget is more specific than the location in the chunk. In the second example, the nugget *[[A United Nations policeman has been shot dead]]* also provides more specific information than the chunk and the citation. We capture the difference in information content of nugget and chunk by assigning the nugget a degree of membership in the chunk (cDM) that is less than 1.

The degree of support (DS), on the other hand, in both examples is 1, since information in the chunks is fully supported by the citations.

The citation metrics, as discussed below, take into account not just the degree of support the citations provide for chunks, but also the degree of membership of nuggets in chunks.

Nuggets with a chunk degree of membership (cDM) of either 0 or 1 are assigned their membership values automatically. If a nugget’s focus window is wholly contained within a chunk, the chunk degree of membership is equal to 1. If no part of the nugget’s focus window occurs in the chunk, then the degree of membership is equal to 0:

Nugget: *A United Nations policeman has been shot dead [[in a sniper attack]] north of the capital Pristina*
Chunk*: policeman has been shot dead ... north of the capital Pristina*
cDM = 0

Nugget: *A United Nations policeman has been shot dead in a sniper attack [[north of the capital Pristina]]*
Chunk: *policeman has been shot dead ... north of the capital Pristina*
cDM = 1

If a nugget is partially contained in a chunk, the chunk degree of membership is manually annotated. Whereas in most cases annotators assign chunk degrees of membership less than 1, for some nuggets, nuggetization can be revisited and a nugget can be broken down into smaller relevant nuggets, as illustrated by the example below.

Nugget: *The UN police said the officer was killed late Sunday [[on the motorway between Leposavic and Mitrovica, some 55 kilometers north of the capital Pristina]]*
Chunk*: the officer was killed … some 55 kilometers north of the capital Pristina*
Citation*: A United Nations policeman has been shot dead in a sniper attack 55 kilometers north of the capital Pristina*

Annotators could revisit nuggetization in this case and break down this nugget into two smaller nuggets, as shown below.

Nugget *1. The UN police said the officer was killed late Sunday [[on the motorway between Leposavic and Mitrovica]], some 55 kilometers north of the capital Pristina*
Chunk*: the officer was killed … some 55 kilometers north of the capital Pristina (cDM=0)*

Nugget 2. *The UN police said the officer was killed late Sunday on the motorway between Leposavic and Mitrovica, [[some 55 kilometers north of the capital Pristina]]*
Chunk*: the officer was killed … some 55 kilometers north of the capital Pristina (cDM=1)*

Degrees of support are intended to estimate to what extent a given citation supports the information in the nugget that is contained in the chunk. If citation does not fully support the part of the chunk which corresponds to a nugget, then a partial degree of support is being assigned, as illustrated in the example below.

Nugget: *Menon was killed [[in northern Kosovo]]*
Chunk: *in … northern Kosovo (cDM=1)*
Citation*: the attack fell last night in Kosovo and that led to the death of one of United Nations policemen (DS=0.5)*

Nugget: *[[A United Nations policeman has been shot dead]] in a sniper attack north of the capital Pristina*
Chunk: *A United Nations policeman has been shot dead … north of the capital Pristina (cDM=1)*
Citation*: The officer was killed between Leposavic and Mitrovica some 55 kilometers north of the capital Pristina (DS=0.8)*

4.2 Unsupported Nugs

Whereas the main goal of the citation checking task is to compute document citation metrics, the output of this task also helped to verify that information in the nugs accurately reflects information in the source documents. Using the output of the citation checking task, we were able to automatically identify all nugs that were not supported by the corpus. Unsupported nugs are usually the result of incorrect machine translations. For example, a foreign source document might have reported that 300 people visited a town, but a machine translated snippet would incorrectly say that people visited 300 towns. If this information were not verified, it would become part of the answer key, incorrectly penalizing other systems. In order to avoid this, as a post-processing step, we reviewed all nuggets that were not fully supported by their citations (i.e. the degrees of support for all citations were less than 1) and modified their relevancy scores. The citation checking task, therefore, allowed us to confirm that no credit is given for information which is not supported by the corpus.

4.3 Missing Context tag

During evaluation, annotators were also asked to use some ‘bookkeeping’ tags, including ‘Missing Context’. The Missing Context tag was assigned when snippets did not provide sufficient context and annotators had to consult the original document to check whether information in the snippet is indeed relevant. For example, the following query has an activity date restriction, but the reported snippet did not provide any dates:

Query: Where has Robert B. Zoellick been between 11/1/ 2005 – 11/30/ 2005
Snippet: Robert Zoellick was in China

In order to make a decision on whether this visit took place in the time period provided by the query, annotators had to consult the documents provided in support of this snippet.

In Phase 2, the Missing Context tag was used in about 30% of all snippets, which means that in 30% of the cases annotators were not able to evaluate the relevancy of the snippets without verifying their citations. The citation checking task, therefore, was necessary not only to identify unsupported information and compute document citation metrics, but also to evaluate the relevance of the reported snippets.

5. Citation Checking Metrics

Given the definitions of chunk degree of membership (cDM) and degree of support (DS) in section 4, we defined metrics for document citation. We gave precise definitions for three key metrics: document recall, document precision, and document F-value. A key technical contribution here is defining these metrics in the presence of multiple, heterogeneous sources of uncertainty and ambiguity.

The building blocks for these metrics are fuzzy measures of the correctness or incorrectness of a document citation for a nugget. A non-fuzzy document metric is based on a simple count of the number of *right*, *wrong*, and *missing* document citations, where these categories are sharp – i.e. a document is either a correct citation or not. Instead, we employ the notions of chunk degree of membership and degree of support, defined in Section 4.1, to define fuzzy document metrics.

To compute document citation metrics for a distiller *A*, we consider each nugget *k* from *A* and each document *d* cited by *A* in support of a chunk containing *k*. Let C_k be the chunk degree of membership of *k* and S_{dk} the degree of support provided by document *d* for the chunk associated with nugget *k*. Then we define the following metrics:

$$\#\text{D-Right}_{\text{A}} = \sum_k \sum_d C_k S_{dk} \qquad (1)$$

$$\#\text{D-Wrong}_{\text{A}} = \sum_k \sum_d C_k (1 - S_{dk}) \qquad (2)$$

Note that while the summation over *k* ranges over all nuggets from distiller *A*, the summation over *d* is confined to documents cited by *A* in support of a chunk containing nugget *k*.

The number of “right” citations for a distiller is therefore a sum of fuzzy measures of document support for each of the distiller’s nuggets. Similarly, the number of “wrong” citations sums over fuzzy values for incorrect citations.

We also need a way of measuring the extent to which the distiller *misses* relevant documents for a nugget. Since we do not have ground truth regarding which documents are relevant to which nuggets, we look at documents cited by *other* distillers for nuggets within the same nug. Suppose, for example, that distiller *A* has a nugget N_A in the nug **NG**. Distiller *B* has one or more nuggets in **NG**. We find the nugget N_B from *B* in that has maximal degree of membership in **NG**. If *B* cites document *D* in support of a chunk that wholly or partially contains N_B while A fails to cite D in support of any chunk wholly or partially containing N_A, then the number of misses attributed to A should increase to the degree that D is a correct citation. But this is just the degree to which D counts in favor of B's #Right score. In other words, distiller A's misses are the citations other distillers got right but A didn't.

Since the same document may be cited by multiple distillers, we define a distiller independent degree of rightness, *r(d)*, for document *d* by:

$$r(d) = \max_k \{C_k S_{dk}\} \qquad (3)$$

where *k* now ranges over all nuggets from any distiller.

Let M_A be the set of documents cited by some distiller but not cited by distiller *A* (the set of missing documents for *A*). Then we define

$$\text{\#D-Missing}_\text{A} = \sum_{d \in M_A} r(d) \qquad (4)$$

We can now define these document metrics for distiller *A*:

$$\text{D-Recall}_\text{A} = \frac{\text{\#D-Right}_\text{A}}{\text{\#D-Right}_\text{A} + \text{\#D-Missing}_\text{A}} \qquad (5)$$

$$\text{D-Precision}_\text{A} = \frac{\text{\#D-Right}_\text{A}}{\text{\#D-Right}_\text{A} + \text{\#D-Wrong}_\text{A}} \qquad (6)$$

$$\text{D-F-Value}_\text{A} = \frac{\text{2D-Precision}_\text{A}(\text{D-Recall}_\text{A})}{\text{D-Precision}_\text{A} + \text{D-Recall}_\text{A}} \qquad (7)$$

6. Information Content Metrics

Recall that the overall goal of distillation is to identify relevant, nonredundant information, and to provide all citations that support this information. To explain how the citation metrics were used to measure overall distillation performance and what their impact was in the evaluation, we need to describe a different set of metrics with which the citation metrics interacted. These are metrics that measure the performance of the distillers in providing relevant *information* in response to queries, as contrasted to citation metrics, which measure how well distillers do in supporting extracted information with document citations.

The information content metrics are based on two key factors: (1) the degree of relevance of *nugs* (not nuggets) to the query; and (2) the degree of membership of nuggets in nugs. Recall that a nug is a collection of nuggets that have similar information content and that may differ with regard to specificity with which they capture that content. The information content of a nug is the information content of the most specific nugget in the nug. Nugs are annotated with a number between 0 and 1 indicating their *degree of relevance* to the query. Nuggets are assigned a *degree of membership* in nugs, with the most specific nugget in a nug having degree of membership 1 and less specific nuggets in the nug having degrees of membership less than 1. These fuzzy measures are used to define information content precision, recall, and F-value.

Suppose that we are dealing with distiller *A*. We use the following formulas, in which summations extend over all of the *nugs* (not nuggets) 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):

$$\text{\#I-Right}_\text{A} = \sum_k R_k D_k \qquad (8)$$

$$\text{\#I-Wrong}_\text{A} = \text{EW}_\text{A} + (\sum_k (1 - R_k) D_k + \text{\# Redundant}_\text{A}(k)) \qquad (9)$$

$$\text{\#I-Missing}_\text{A} = \sum_k R_k (1 - D_k) \qquad (10)$$

In (9), EW_A is an estimate of the number of wrong nuggets in un-nuggetized text. *#Redundant$_A$(k)* is the number of redundant nuggets distiller *A* has in nug *k*. A redundant response (one that repeats information already provided) was regarded as incorrect for the purposes of scoring information retrieval.

Using the counts (8)-(10), the metrics I-Recall$_\text{A}$, I-Precision$_\text{A}$, and I-F-Value$_\text{A}$ are defined in the same way as the corresponding document metrics are defined.

In order to incorporate citation strength in the information content score, we weighted the information recall score for each distiller by the *document F-value.* The rationale for doing so is that retrieved information not backed by citations is of questionable value and this should be reflected in the information recall score. The weighting is done through the formula:

$$\text{CW-Recall} = \frac{1}{N} \sum_{n=1}^{N} \bar{D}_n \sqrt{\text{D-F-value}_\text{n}} \qquad (11)$$

In (11), the summation is over all nugs; $\bar{D}_n$ is the mean degree of membership of the distiller's nuggets in nug *n*, and *D-F-value$_n$* is the *D-F*-value for the document citations provided by the distiller for the nuggets in nug *n*. We used the square root of *D-F-value$_n$* to soften the impact of low citation *F*-values on citation-weighted recall.

This weighted recall was used in place of normal recall in the formula for information F-value to yield what we called *citation-weighted F-value*, which is defined as

$$\text{CW-F-Value} = \frac{\text{2I-Precision(CW-Recall)}}{\text{I-Precision + CW-Recall}} \quad (12)$$

7. The Results

The final results compared system and human performance. As expected, systems were able to analyze a much larger collection of documents than humans, and they generate high recall scores. The following table shows the average ratios of system scores over human scores for four genres: Structured text (newswire), unstructured text (blogs and newsgroups), structured audio (broadcast news) and unstructured audio (broadcast conversation). ("CW" below stands for "citation-weighted".) For each query, there were either two or three human distillers and their scores were averaged in computing the ratio of machine to human performance.

Average	Info. F-value ratio	Doc Recall ratio	Doc Precision ratio	Doc F-value ratio	CW F-value ratio
SText	0.51	0.88	0.98	0.94	0.54
UText	0.78	1.41	1.08	1.31	0.80
SAudio	0.50	0.78	1.11	0.84	0.53
UAudio	0.28	0.34	1.03	0.38	0.26

Table 1: Average ratios of machine to human information and document scores by source type.

As Table 1 shows, machine distillers outperformed humans on document recall for unstructured text. Machine distillers also did marginally better than humans on document precision in that category. The category that gave machine distillers the most difficulty as compared to humans was unstructured audio, where document recall was on average only a third of that of humans. In all other cases, the document F-value for the machine distillers was high enough that the citation weighted F-value for machine distillers was high than the simple information F-value – i.e. machine distillers benefited from taking into account citation metrics in the final score.

Average	Info. F-val ratio	Doc Recall ratio	Doc Precision ratio	Doc F-val ratio	CW F-val ratio
English	0.61	0.70	1.02	0.75	0.59
Chinese	0.35	0.62	1.08	0.63	0.36
Arabic	0.53	1.30	1.05	1.28	0.61

Table 2. Average ratios of machine to human information and document scores by source language.

When we consider the results by language (Table 2), we see that for English and Chinese, machine distillers' document recall was lower than that of humans, but document recall for Arabic exceeded that of humans by 30%. Although we would not have expected this *a priori*, document precision for machine distillers exceeded that of the humans for all three languages. Overall, taking into account document F-value helped the machine distillers in two out of three languages.

A more detailed analysis shows that for the 39 combinations of source type and distiller, only one final score was decreased by using CW F-value instead of information F-value, whereas eight were increased. In general, therefore, we can say that the systems benefited from including document citation metrics in the evaluation.

8. Conclusion

We have described the approach to the evaluation of document citations used for the DARPA GALE Phase 2 Distillation evaluation. This approach is a principled combination of degrees of support and a measure of information content overlap. We showed that the citation checking task is necessary not only to compute document citation metrics, but also to verify relevance of system responses, as well as to identify unsupported information. We also showed how citation metrics may be combined with information metrics to yield an overall measure of the performance of a distillation system in answering queries. A noteworthy implication of our evaluation of machine and human distillers is that the performance of machines relative to humans is on average better when document retrieval is taken into account than when information content alone is considered in evaluating performance.

9. Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency DARPA/IPTO, Global Autonomous Language Exploitation, 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

10. References

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

Voorhees, E. 2003. Overview of the TREC 2003 question *answering* track. In Proceedings of TREC 2003.

White J.V, D. Hunter, O. Babko-Malaya, C. Fournelle, M. K. Schneider "Evaluation of Redundant Information from Distillation Systems using Nuggets and Fuzzy Sets" in Proceedings of SIGIR 2009 Workshop Redundancy, Diversity, and Interdependent Document Relevance (IDR '09)

White J. V., Hunter D., Goldstein J.D. 2008 "Statistical Evaluation of Information Distillation Systems", in *Proceedings* of LREC 2008.