Annotation of Nuggets and Relevance in GALE Distillation Evaluation[1]

Olga Babko-Malaya
BAE Systems Advanced Information Technologies
6 New England Executive Park, Burlington, MA, 01803, US
Olga.Babko-Malaya@baesystems.com

Abstract

This paper presents an approach to annotation that BAE Systems has employed in the DARPA GALE Phase 2 Distillation evaluation. The purpose of the GALE Distillation evaluation is to quantify the amount of relevant and non-redundant information a distillation engine is able to produce in response to a specific, formatted query; 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. As part of the evaluation, following NIST evaluation methodology of complex question answering (Voorhees, 2003), human annotators were asked to establish the relevancy of responses as well as the presence of atomic facts or information units, called nuggets of information. This paper discusses various challenges to the annotation of nuggets, called nuggetization, which include interaction between the granularity of nuggets and relevancy of these nuggets to the query in question. The approach proposed in the paper views nuggetization as a procedural task and allows annotators to revisit nuggetization based on the requirements imposed by the relevancy guidelines defined with a specific end-user in mind. This approach is shown in the paper to produce consistent annotations with high inter-annotator agreement scores.

1. Introduction

Quantitative evaluations of question answering, such as TREC QA evaluations conducted by NIST employ a notion of an “information nugget” for evaluating answers to complex questions (Voorhees, 2003). Unlike factoid questions, such as “Where is Rider College located?”, which can be answered by simply extracting named entities (persons, organization, locations, dates, etc) from documents, answers to the so called “definition”, “relationship” and “opinion” questions are unstructured and open ended. To assess the quality of these open-ended answers, NIST developed an evaluation methodology in which humans are asked to establish the presence of important facts, or nuggets of information, in system responses. Several papers, on the other hand, have raised the question of whether human-based nugget annotations are stable and whether it is possible to define the appropriate granularity level for nuggets (e.g. Lin and Zhang, 2007). The answers have important implications for the reliability of system scores.

The paper presents an approach to nuggetization which has been employed by BAE Systems Advanced Information Technologies in the DARPA GALE Phase 2 Distillation evaluation. The goal of this evaluation is to compare the distillation performance of bilingual humans using non-GALE state-of-the-art search tools to the performance of GALE distillation engines run directly by research developers. In response to a certain query, distillers, whether they are humans or machines, produce responses consisting of snippets of text, and as part of the evaluation, human annotators “nuggetize” the responses, i.e. identify relevant nuggets of information.

As opposed to “computationally straightforward” approaches to nuggets (e.g. Lin and Demner-Fushman, 2005; Marton and Radul, 2006; Zhou and Hovy, 2007), nuggetization in GALE is defined “procedurally”, based on a small set of predefined rules. Whereas some of these rules are linguistic in nature, others are based on “relevance” or importance of created nuggets. The approach to nuggetization discussed in the paper has two goals. First, human annotators should be able to apply nuggetization rules consistently and reliably, with high inter-annotator agreement. Second, all nuggets must be relevant to the query in question, where relevancy depends on the needs of the end user of distillation systems. This interaction between nuggetization and assessing relevance of nuggets presents a big challenge for Distillation evaluation given that judging relevance is inherently subjective (Shamber, 1994 and Voorhees, 2003).

The approach to nuggetization adopted in the paper views nuggetization and relevance as separate tasks, where annotators first use nuggetization rules to define nuggets of fixed granularity, and later can revisit these nuggets based on the requirements imposed by the relevancy guidelines defined with a specific end-user in mind. This approach, as shown in the paper, has proven to produce consistent annotations with high inter-annotator agreement scores.

2. GALE Distillation Evaluation Overview

The purpose of the GALE Distillation evaluation is to quantify the amount of relevant and non-redundant information a distillation engine is able to produce in response to a specific, formatted query, and to compare that amount of information to the amount of information gathered by a bilingual human using commonly available

[1] 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.

state-of-the-art tools, such Copernic or Onfolio. GALE engines distill data from audio and text sources in multiple languages and produce English-only responses using translations and transcriptions. These responses are evaluated in two separate dimensions, information content and document support.

The queries conform to templates, which contain argument variables that range over events, topics, people, organizations, locations, and dates. In year 2 of the project, 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, which may consist of exact text extractions, translations, summarizations, or paraphrases of the source material. These output responses should be clean and precise: irrelevant and redundant information is penalized. During evaluation, annotators create nuggets out of relevant snippet text, and map them to equivalence classes, called nugs. Nugs contain semantically equivalent nuggets from different distillers and with different source documents. The number of nugs in the irrelevant text is estimated. Whereas recall is the ratio between the number of 'right' (i.e. relevant and non-redundant nugs) and the total number of relevant nugs for a given query, precision is calculated as the ratio between the number of right nugs and the number of all nugs retrieved by a given distiller (see White et al, 2008 for discussion of scoring information content in GALE Distillation Evaluation). Unlike TREC evaluations, where an answer key was created by using responses as well as research performed during the original development of the question, GALE Distillation evaluation is only using the pool of responses produced by machine and human distillers.

After receiving all distiller responses, BAE evaluated each query individually. The first 500 words of each response were manually parsed into nuggets, which are atomic facts or information units that answer the query in question. The nuggets, in turn, were mapped into nugs, which are equivalence classes of one or more nuggets. In Phase 2, BAE also performed citation checking to validate that the content of each nugget accurately reflects information in the source document.

3. Nuggetization Procedure

Scoring of responses during evaluation is affected by nuggetization in two ways. First, all characters which are not included in at least one nugget count as irrelevant (thus reducing precision score); nuggets, therefore, are annotated to include the maximal extent of relevant information. Identifying maximal extent, however, is not always easy, given that in many cases responses include background or other contextual information, and humans often disagree on where the boundary is between "relevant" and "irrelevant" text.

Second, the number of nuggets created from "relevant text" has a significant impact on both precision and recall scores. For each nugget created out of a distiller's response, the distiller gets credit for providing an "on-target" piece of information, whereas other distillers are penalized for missing this information. The granularity of nuggets, therefore, is critical for comparing systems' performance.

To address both of these issues, we have created two sets of nuggetization rules. The first set of rules is used to define nuggets with a fixed granularity. Whereas some of these rules might be viewed as arbitrary, the goal of this task is to set up an annotation procedure which could be easily followed by human annotators. The second set of rules, on the other hand, allows annotators to revisit their nuggetization strategy, where initial granularity of nuggets might be changed to incorporate the relevancy requirements, or as the result of the other annotation tasks.

3.1 Nuggetization Rules

Nuggetization is a process of breaking down a snippet into 'atomic' facts or units of information. For example, consider the following snippet:

Before receiving the war spending bill from Congress, the president will fly to the headquarters of the U.S. Central Command in Tampa, Florida

There is a large number of ways to break this sentence down into smaller pieces of information. The main question, however, is which level of granularity would correspond to the notion of an 'atomic' fact.

In order to ensure consistent annotations, we have not been trying to achieve the finest level of granularity, but rather developed a set of procedural rules listed below. Whereas some nuggetization rules rely on linguistic criteria, others are specified based on the end-user specified significance.

The first nuggetization rule says that

- Nuggets are created out of each core verb and its arguments, where the maximal extent of the argument is always selected.

Since there are two core verbs in the sentence above, two nuggets are generated. The extent of the nugget in each example is indicated by double brackets:

Nugget 1. Before receiving the war spending bill from Congress, [[the president will fly to the headquarters of the U.S. Central Command]] in Tampa, Florida

Nugget 2. Before [[receiving the war spending bill from Congress, the president]] will fly to the headquarters of the U.S. Central Command in Tampa, Florida

The second rule says that

- All temporal, locative, causative and other types of modifiers of the verb constitute a separate nugget, including subordinate clauses, where the maximal extent of the modifier is always selected

In this example, a nugget is created out of the temporal clause:

Nugget 3: [[Before receiving the war spending bill from Congress]], the president will fly to the headquarters of the U.S. Central Command in Tampa, Florida

It is certainly possible to break this snippet down into much smaller pieces, for example, by breaking down noun phrases into *'the war spending bill'* and *'bill from Congress'*. As we will see in the next subsection, however, this level of granularity is often too fine grained for the task in question, where nuggets also serve as relevant answers to the Distillation queries. In order to provide the level of granularity which corresponds to possible answers to the queries, as well as to simplify the annotation task, the following rule was introduced:

- Noun phrases are not decomposed into separate nuggets, unless they contain temporal, locative, numerical information, or titles.

These exceptions relate to the "significance" of created nuggets, where times and locations of events, as well as quantitative information are considered to be important pieces of information. Based on this rule, only one other nugget is being created for the snippet above, which specifies the location of the headquarters of the U.S. Central Command:

Nugget 4: Before receiving the war spending bill from Congress, the president will fly to the headquarters of the U.S. Central Command [[in Tampa, Florida]]

Given these rules, temporal, locative, numerical expressions and titles always constitute a nugget.

For example, given the snippet *The five attacks resulted in 80 deaths*, 3 nuggets are being created:

Nugget 1: The five [[attacks resulted in]] 80 [[deaths]]
Nugget 2: [[The five]] attacks resulted in 80 deaths
Nugget 3: The five attacks resulted in [[80]] deaths

Times and locations make a nugget independent on whether they modify a verb or a noun:

Snippet: A U.N. policeman from India's southern state of Kerala, was shot to death

Nugget 1: [[A U.N. policeman]] from India's southern state of Kerala, [[was shot to death]]

Nugget 2: A U.N. policeman [[from India's southern state of Kerala]], was shot to death

Nuggets created for titles include both the predicate and its argument. For example, the snippet *Israeli Prime Minister Ariel Sharon talked with Ben-Ami* yields two nuggets, the first one focuses on the fact that Sharon was Israeli Prime Minister, and the second one on the core clause:

Nugget 1: [[Israeli Prime Minister Ariel Sharon]] talked with Ben-Ami.
Nugget 2: [[Israeli Prime Minister Ariel Sharon talked with Ben-Ami.]]

Special rules are defined for certain syntactic constructions. For example, a separate rule is defined for conjoined phrases, which states that

- When possible, conjuncts should be broken down into separate nuggets.

For example, the snippet *Iraq can import food, medicines and other goods needed for the country's shattered Infrastructure* yields 3 nuggets:

Nugget 1: [[Iraq can import food]], medicines and other goods needed for the country's shattered infrastructure
Nugget 2: [[Iraq can import]] food, [[medicines]] and other goods needed for the country's shattered infrastructure
Nugget 3: [[Iraq can import]] food, medicines [[and other goods needed for the country's shattered infrastructure]]

Note that not all conjuncts can be split into several nuggets. For example, [[*John and Mary met*]] is one nugget, rather than two.

And, finally, special rules are created for sentences with direct and indirect quotations:

- When creating nuggets with the verb of saying as being the core of the nugget, direct quotations are not decomposed into nuggets.

[[``So we have to infringe on Freeman's religious beliefs because of what someone else might do,'' ACLU legal director Randall C. Marshall said.]]

- In the case of indirect speech, utterances are not broken down into smaller nuggets unless the utterance has conjoined clauses:

Nugget 1: [[Iraq's Deputy Prime Minister Tariq Aziz said that ballistic missiles held by Baghdad do not violate UN accords,]] and Iraq would welcome more UN weapons inspectors in the country.

Nugget 2: [[Iraq's Deputy Prime Minister Tariq Aziz said]] that ballistic missiles held by Baghdad do not *violate UN accords, [[and Iraq would welcome more UN weapons inspectors in the country]].*

Note that in sentences with direct and indirect quotations, two types of nuggets can be generated. For example, the snippet *A spokeswoman said that there were no details available* generates two nuggets, the first one focuses on the content of the utterance, and the second one on the statement made by the spokeswoman:

Nugget1: A spokeswoman said that [[there were no details available]]
Nugget2: [[A spokeswoman said that there were no details available]]

The granularity of nuggets defined by these rules is certainly not atomic, in the sense that many of these nuggets can be viewed as containing more than one fact. On the other hand, the purpose of this procedure is to define a level of granularity which (1) could be easily followed by human annotators and (2) would correspond as closely as possible to the basic relevant answers to the queries in GALE Distillation.

3.2 Revisiting Nuggetization

Whereas the set of rules above requires annotators to create nuggets with a fixed level of granularity, nuggetization can sometimes be revisited. As discussion below shows, nuggets can be revisited in the following situations:

- a nugget is too fine grained to serve as a relevant answer to the query
- a nugget cannot be interpreted as a fact inferred from the snippet
- the granularity of a nugget is different from granularity of nuggets in semantically equivalent classes, or nugs
- a nugget is only partially supported by the citations.

3.2.1. Nuggetization and Relevancy

The main reason for revisiting granularity of nuggets is based on the following rule: “Do not break down relevant text into nuggets if these nuggets are not relevant on their own”. For example, consider the query: DESCRIBE THE RELATIONSHIP BETWEEN [Mitt Romney] AND [Thomas Finneran]. The following text provides a relevant answer to the query: *Finneran invited Mitt Romney to a baseball game he could not attend because of a previous engagement.* If this text is nuggetized according to the nuggetization rules above, 3 nuggets should be created:

Nugget1:[[Finneran invited Mitt Romney to a baseball game]] he could not attend because of a previous engagement
Nugget2: Finneran invited Mitt Romney to a baseball game [[he could not attend]] because of a previous engagement
Nugget3: Finneran invited Mitt Romney to a baseball game he could not attend [[because of a previous engagement]]

However, the facts that *Mitt Romney could not attend a baseball game* and that *he had a previous engagement* do not really answer this query, which asks about the relationship between Mitt Romney and Thomas Finneran. In cases like that, when a nugget is not on-target when considered on its own, but is relevant when considered as part of a larger piece of text, one large nugget can be created: *[[Finneran invited Mitt Romney to a baseball game he could not attend because of a previous engagement.]]*

It is certainly not surprising that granularity of nuggets depends on the template of the query: whereas answers to the queries from the relationship template shown above rarely contain nuggets smaller than clauses or sentences, queries which ask about acquaintances of a person could be answered with nuggets which contain just a person name:

Query: FIND ACQUAINTANCES OF [George W. Bush]
Nugget: George W. Bush telephoned [[Kirchner]] to wish him well

In order to improve agreement scores on relevancy judgments, we have developed relevancy guidelines. For each template, these guidelines specify (1) description of information which should be viewed as relevant, and (2) a list of relevant categories, which characterize the type of relevant information. For example, the relevant categories for the template LIST FACTS ABOUT [event] are as follows:

- time, location, cause/intention/planning, participant, subevent/execution/manner, consequence/reaction/significance

While nuggetizing, annotators have to categorize each nugget with these template-specific categories, which helps them in making decisions about relevancy of created nuggets:

Query: LIST FACTS ABOUT [Vice President Dick Cheney's shooting of Harry Whittington]
Nuggets:
SUBEVENT/EXECUTION/MANNER: [[Whittington was shot]] with pellets
SUBEVENT/EXECUTION/MANNER: Whittington was shot [[with pellets]]
LOCATION: [[on a private Texas ranch]]
MANNER: [[Shooting was accidental]]
REACTION: [[Shooting has generated much talk]]

Some of the relevancy rules, on the other hand, are general and apply across all templates. For example, we have been assuming that for all templates, if an event is relevant to the query, then the time and location of that event is also relevant. Nuggets created by these rules have secondary tags attached to the main relevancy categories. In the following example, the snippet describes a consequence/reaction of the event and says that this consequence happened at a certain location. Separate nuggets are then created for the consequence, as well as its location:

Query: LIST FACTS ABOUT EVENT [The Hamas victory in Palestinian parliamentary elections]
Nuggets:
CONSEQUENCE: [[Russia invited Hamas leaders]] to Moscow [[for talks]]
CONSEQUENCE-LOC: Russia invited Hamas leaders [[to Moscow]] for talks

Relevancy categories and rules discussed above helped to improve inter-annotator consistency for this difficult task, where granularity of nuggets depends on its relevancy with respect to a given query, but relevancy also depends on the template.

3.2.2. Interpretation of nuggets

The granularity of nuggets can also be revisited if the nuggets cannot be interpreted as facts inferred from the snippet. The process of nuggetization is the process of breaking down information in the snippet into atomic pieces of information. All nuggets, therefore, are facts, which (1) can be paraphrased as a sentence, and (2) are inferred from the meaning of the snippet.

Some examples of how nuggets can be paraphrased as a sentence are shown below:

[[British Foreign Secretary Jack Straw]]: Jack Straw is the British Foreign Secretary
Smith visited China [[last Sunday]]: The time of Smith's visit to China is last Sunday
[[One]] person was killed: the number of people killed is one
[[Matish Menon, 34]]: Matish Menon is 34

However, the snippet *Sales exceeded last year's peak* is not broken down into two nuggets, as required by the rules above, since they cannot be paraphrased as informative facts inferred from the snippet.

Nugget: [[Sales exceeded last year's peak]]
NOT: [[Sales exceeded]] last year's [[peak]]
Sales exceeded [[last year's]] peak

The requirement that all nuggets are facts which must be inferred from the meaning of the snippet can also be illustrated by the example below, where nuggets are not created for the following infinitival verbs:

NOT: The court asked [[the prosecution to ask the CIA to reveal blacked-out portions of its cables]]
The court asked the prosecution to ask [[the CIA to reveal blacked-out portions of its cables]]

The snippet does not assert that CIA revealed blacked-out portions of the cables, nor does it assert that the prosecution asked the CIA to reveal them. The only nugget which is created in this case is the one that contains the whole sentence:

Nugget: [[The court asked the prosecution to ask the CIA to reveal blacked-out portions of its cables]]

3.2.3. Merging nuggets into semantically equivalent classes

Granularity of nuggets can also be revisited based on the other annotation tasks. One of such tasks is merging nuggets into semantically equivalent classes of nuggets, called nugs. Nugs are needed on order to compute the recall scores for each distiller, as well as to identify redundant information (see White et al, 2008 for discussion of metrics and scoring).

The task of clustering nuggets into equivalent pieces of information occasionally required revisiting nuggetization. Some nuggets could not be merged into the same nug if they had different granularity. In the example below, snippet 2 is initially broken down into two nuggets: n2 and n3, given that n2 "*Under the oil for food program*" is a modifier in this sentence, rather than an argument. In snippet1, "*the oil for food program*" serves as an argument of the verb, and therefore only one nugget n1 is created. While trying to merge these nuggets, annotators were allowed to revisit their initial nuggetization, and merge nuggets n2 and n3 into one nugget *[[Under the oil for food program, Iraq is allowed to sell oil]]*, which is equivalent to n1.

Snippet 1: *The oil for food program allowed Iraq to sell oil*
Nugget n1: [[The oil for food program allowed Iraq to sell oil]]

Snippet 2: *Under the oil for food program, Iraq is allowed to sell oil*
Nugget n2: [[Under the oil for food program,]] Iraq is allowed to sell oil
Nugget n3: Under the oil for food program, [[Iraq is allowed to sell oil]]

3.2.4. Citation Checking

Another annotation task which might require revisiting nuggetization is citation checking.

The goal of the citation checking task is to evaluate the quantity and relevancy of citations provided by a distiller in support of a given snippet. As part of this task, human annotators had to verify that relevant information in the snippet is indeed supported by the citations provided by the distiller. As mentioned above, snippets could include direct quotations from source text, transcripts or translations of foreign materials, and summarizations of these materials. In addition to snippets, however, machine distillers were also required to submit 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 (interpretable independent of any other resource) and may contain non-contiguous text strings.

- *Citations,* which refer to specific snippet chunks. Each citation indicates the chunk it supports, and includes the literal excerpt from the source document from which it originates (text string in the source language or audio time stamp).

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

Snippet: The UN police said the officer was killed late Sunday on the motorway between Leposavic and Mitrovica, some 55 kilometers (33 miles) north of the capital Pristina

Chunk 1: The UN police said the officer was killed late Sunday on the motorway between Leposavic and Mitrovica

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 between Leposavic and Mitrovica

Chunk 2: The officer was killed late Sunday some 55 kilometers (33 miles) 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 (33 miles) north of the capital, Pristina, police said

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

In order to verify that information in the snippet is supported by the citations, annotators are asked to check for each of the nuggets, if information in the chunk which corresponds to the nugget is indeed supported by the citation.

Occasionally, the granularity of nuggets turned out not to be fine-grained enough for the citation checking task. For example, based on the ‘maximal extent’ rule, an annotator first created the following nugget for the locative modifier: *The UN police said the officer was killed late Sunday [[on the motorway between Leposavic and Mitrovica, some 55 kilometers (33 miles) north of the capital Pristina]]*

During the citation checking task, annotators are not able to confirm that this nugget is supported by both citations, given that both citations only partially support this nugget: whereas citation 1 verifies that the event took place “*on the motorway between Leposavic and Mitrovica*”, citation 2 only confirms that it took place “*some 55 kilometers (33 miles) north of the capital Pristina*”.

Given that this nugget can be easily broken down into smaller relevant nuggets, annotators have to revisit nuggetization and split this locative modifier into 2 nuggets, so that full credit can be given to the distiller for providing correct chunks and citations:

The UN police said the officer was killed late Sunday [[on the motorway between Leposavic and Mitrovica]], some 55 kilometers (33 miles) north of the capital Pristina

The UN police said the officer was killed late Sunday on the motorway between Leposavic and Mitrovica, [[some 55 kilometers (33 miles) north of the capital Pristina]]

4. Nuggetization and World Knowledge

While annotating relevancy, we often noted that a decision on whether a certain snippet is relevant or not depends on annotator’s knowledge about the facts of the world. For example, in the following example, an annotator would judge the snippet as relevant only if she knew that Dr. Ayman Zawahiri was closely related to the Egyptian Islamic Jihad (as the second and the last “emir”).

Query: PROVIDE INFORMATION ON [the Egyptian Islamic Jihad]
Snippet: Dr. Ayman Zawahiri's decade-long quest to weaponize and use anthrax against US targets

Reasoning and world knowledge is not part of the Distillation task, and Distillation engines were not required to provide relevant information when relevancy depends on the knowledge of the facts of the world. On the other hand, the NLP tools used by distillation engines were often able to extract information of this type. Surprisingly, distillation engines were even better in finding relevant information linked by co-reference than human distillers.

Annotation of such snippets as relevant or irrelevant affects scoring. If this snippet were judged relevant, then the recall of the system which did not find it would be reduced for missing this information (see White et al, 2008). If this snippet were judged as irrelevant, then the precision of the systems which found this information would be reduced for providing irrelevant information. Both situations are clearly wrong: whereas we do not want to penalize the team who did not find this snippet, we certainly should not penalize the team who was able to find it.

To solve this problem, we have introduced a special tag to mark ‘world knowledge’ cases. Nuggets were marked with this tag, if world knowledge or reasoning was needed in order to judge whether the nugget was relevant or not. This tag was further used to solve the scoring problem: whereas the distiller who found this snippet received credit for this information, the recall of the distillers who did not find it was not reduced for missing this information.

5. Inter-Annotator Agreement

In order to ensure reliable and consistent annotations, each query was at least double annotated and adjudicated. The following metrics show inter-annotator agreement scores.

Relevancy Agreement reflects yes/no judgments on a snippet level. Every annotated snippet that has at least one nugget is considered relevant. Snippets from which no nuggets were created are irrelevant. Relevancy Agreement counts the percentage of snippets which both annotators judged as relevant or irrelevant.

Nugget Overlap. Every relevant snippet is associated with a set of nuggets, which are represented as intervals over the text of the snippet. The Overlap and difference (Diff) between two annotators’ nugget intervals are calculated, counting only ‘meaningful characters’, i.e. letters and numbers. The Overlap is the number of meaningful characters which both annotators included in the extent of the nuggets, and Diff is the number of meaningful characters which only one annotator included in the extent of the nuggets. The following formula is then used to determine percentage of nugget overlap:

% NuggetOverlap = Overlap / (0.5*Diff + Overlap)

The table below shows Phase 2 inter-annotator agreement scores across languages and different source types.

	Relevance Agreement	**%Nugget Overlap**
Audio	0.90	0.83
Text	0.89	0.82
Arabic	0.91	0.86
Chinese	0.85	0.72
English	0.91	0.85
Structured	0.89	0.79
Unstructured	0.89	0.80
Total	0.89	0.80

Table 1: Inter-Annotator Agreement Scores

6. Conclusion

The paper presented an approach to nuggetization which has been employed by BAE Systems in the DARPA GALE Phase 2 Distillation evaluation. It discussed various challenges to the annotation of system responses, including the interaction of the granularity of nuggets created by human annotators and relevancy of these nuggets to the query in question. High inter-annotator agreement scores reported in the paper support the current approach and show that this task can be done consistently by human annotators.

7. Acknowledgements

I would like to thank Connie Fournelle, Jim White, Christopher White, and Dan Hunter for discussions and valuable comments, the team of annotators who diligently tried to follow all the rules discussed in this paper, and Gayle McElvain for all her help with this project including computing inter-annotator agreement scores. All remaining errors are mine.

8. References

Lin, J. and D. Demner-Fushman, 2005. Automatically evaluating answers to definition questions. In *Proceedings of the Human Language Technology Conference* (HLT-EMNLP 2005).

Lin J. and P. Zhang. 2007. Deconstructing Nuggets: The Stability and Reliability of Complex Question Answering Evaluation. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007),* pages 327-334, Amsterdam, the Netherlands

Marton, G. and A. Radul, 2006. Nuggeteer: automatic nugget-based evaluation using description and judgments. In *Proceedings of NAACL-HLT* 2006

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

Shamber L. 1994. Relevance and information behavior. In *Annual Review of Information Science and Technology.* 29:3-48.

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

White J.V., D. Hunter, and J.D.Goldstein, 2008. “Statistical Evaluation of Information Distillation Systems”, submitted to LREC 2008