LAMP-TR-127
CS-TR-4771
UMIACS-TR-2005-71

December 2005

Will Pyramids Built of Nuggets Topple Over?

Jimmy Lin[†] and Dina Demner-Fushman[‡]

[†]College of Information Studies
[‡]Department of Computer Science
[†,‡]Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742
jimmylin@umd.edu, demner@cs.umd.edu

Abstract

The present methodology for evaluating complex questions at TREC analyzes answers in terms of facts called "nuggets". The official F-score metric represents the harmonic mean between recall and precision at the nugget level. There is an implicit assumption that some facts are more important than others, which is implemented in a binary split between "vital" and "okay" nuggets. This distinction holds important implications for the TREC scoring model—essentially, systems only receive credit for retrieving vital nuggets—and is a source of evaluation instability. The upshot is that for many questions in the TREC testsets, the median score across all submitted runs is zero. In this work, we introduce a scoring model based on judgments from multiple assessors that captures a more refined notion of nugget importance. We demonstrate on TREC 2003, 2004, and 2005 data that our "nugget pyramids" address many shortcomings of the present methodology, while introducing only minimal additional overhead on the evaluation flow.

Last updated: December 9, 2005
Keywords: question answering, nuggets, complex questions, definition questions, evaluation

1 Introduction

The field of question answering has been moving away from simple "factoid" questions such as "Who invented the paper clip?" to more complex information needs such as "Who is Aaron Copland?" and "How have South American drug cartels been using banks in Liechtenstein to launder money?", which cannot be answered by simple named-entities. Over the past few years, NIST through the TREC QA tracks has implemented an evaluation methodology based on the notion of "information nuggets" to assess the quality of answers to such complex questions. This paradigm has gained widespread acceptance in the research community, and is currently being applied to evaluate answers to so-called "definition", "relationship", and "opinion" questions.

Since quantitative evaluation is arguably the single biggest driver of advances in language technologies, it is important to closely examine the characteristics of a scoring metric to ensure its fairness, reliability, and stability. In this work, we identify a potential source of instability in the nugget evaluation paradigm, develop a new scoring modeling, and demonstrate that our new model addresses some of the shortcomings of the original methodology. Our new variant stands as a proposed metric for evaluating answers to complex questions in the TREC 2006 question answering track.

The paper is organized as follows: Section 2 provides a brief overview of the nugget evaluation methodology. Section 3 draws attention to the vital/okay nugget distinction and the problems it creates. Section 4 outlines our proposal for building "nugget pyramids", a more-refined model of nugget importance that combines judgments from multiple assessors. Section 5 describes our methodology for evaluating this new model, and Section 6 presents our results. A discussion of related issues occupies Section 7, and the paper concludes with Section 8.

2 Evaluation of Complex Questions

To date, NIST has conducted three large-scale evaluations of complex questions using a nugget-based evaluation methodology: "definition" questions in TREC 2003, "other" questions in TREC 2004 and TREC 2005, and "relationship" questions in TREC 2005. Since relatively few teams participated in the 2005 evaluation of "relationship" questions, this work focuses on the three year's worth of "definition/other" questions. The nugget-based paradigm has been previously detailed in a number of papers (Voorhees, 2003; Hildebrandt et al., 2004; Lin and Demner-Fushman, 2005a); here, we present only a short summary.

System responses to complex questions consist of an unordered set of passages. To evaluate answers, NIST pools answer strings from all participants, removes their association with the runs that produced them, and presents them to a human assessor. Using these responses and research performed during the original development of the question, the assessor creates an "answer key" comprised of a list of "nuggets"—essentially, facts about the target. According to TREC guidelines, a nugget is defined as a fact for which the assessor could make a binary decision as to whether a response contained that nugget (Voorhees, 2003). As an example, relevant nuggets for entity "AARP" are shown in Table 1. In addition to creating the nuggets, the assessor also manually classifies each as either *vital* or *okay*. Vital nuggets represent concepts that must be present in a "good" definition; on the other hand, okay nuggets contribute worthwhile information about the target but are not essential. The vital/okay distinction has significant implications, demonstrated below.

Once the answer key of vital/okay nuggets is created, the assessor then goes back and manually scores each run. For each system response, he or she decides whether or not each nugget is present. The final F-score for an answer is calculated in the manner described in Figure 1, and the final score of a system run is the mean of scores across all questions. The per-question F-score is a harmonic mean between nugget precision and nugget recall, where recall is heavily favored (controlled by the β parameter, set to five in 2003 and three in 2004 and 2005). Nugget recall is calculated solely on

vital	30+ million members
okay	Spends heavily on research & education
vital	Largest seniors organization
vital	Largest dues paying organization
vital	Membership eligibility is 50+
okay	Abbreviated name to attract boomers
okay	Most of its work done by volunteers
okay	Receives millions for product endorsements
okay	Receives millions from product endorsements

Table 1: Answer nuggets for the target "AARP"

Let

r	# of *vital* nuggets returned in a response
a	# of *okay* nuggets returned in a response
R	# of *vital* nuggets in the answer key
l	# of non-whitespace characters in the entire answer string

Then

$$\begin{aligned} \text{recall } (\mathcal{R}) &= r/R \\ \text{allowance } (\alpha) &= 100 \times (r + a) \\ \text{precision } (\mathcal{P}) &= \begin{cases} 1 & \text{if } l < \alpha \\ 1 - \frac{l-\alpha}{l} & \text{otherwise} \end{cases} \end{aligned}$$

Finally, the $F_\beta = \dfrac{(\beta^2 + 1) \times \mathcal{P} \times \mathcal{R}}{\beta^2 \times \mathcal{P} + \mathcal{R}}$

$\beta = 5$ in TREC 2003, $\beta = 3$ in TREC 2004, 2005.

Figure 1: Official definition of F-measure.

vital nuggets (which means no credit is given for returning okay nuggets), while nugget precision is approximated by a length allowance given based on the number of both vital and okay nuggets returned. Early in a pilot study, researchers discovered that it was impossible for assessors to enumerate the total set of nuggets contained in a system response (Voorhees, 2003), which corresponds to the denominator in the precision calculation. Thus, a penalty for verbosity serves as a surrogate for precision.

Note that while a question's answer key only needs to be created once, assessors must manually determine if each nugget is present in a system's response. This human involvement has been identified as a bottleneck in the evaluation process, although recently Lin and Demner-Fushman (2005a) have developed an automatic scoring metric called Pourpre that correlates well with human judgments.

3 What's Vital? What's Okay?

Previously, researchers have pointed out the vital/okay distinction as a source of instability in the nugget-based evaluation methodology, especially given the manner in which F-score is calculated (Hildebrandt et al., 2004; Lin and Demner-Fushman, 2005a). Since only vital nuggets figure into the calculation of nugget recall, there is a large "quantization effect" for system scores on topics that have few vital nuggets. For example, on a question that has only one vital nugget, a system cannot obtain a non-zero score unless that vital nugget is retrieved. In reality, whether or not a system returned a passage containing that single vital nugget is often a matter of luck, which is compounded by assessor judgment errors.

The polarizing effect of the vital/okay distinction brings into question the stability of TREC evaluations: in TREC 2003, three question (out of 50) had only one vital nugget and ten had only two vital nuggets; in TREC 2004, two questions (out of 64) had only one vital nugget and fifteen had only two vital nuggets; in TREC 2005, the numbers were five and sixteen (out of 75 questions). For example, "F16" is the target for question 71.7 from TREC 2005. The only vital nugget is "First F16s built in 1974". The practical effect of the vital/okay distinction in its current form is the number of questions for which the median system score across all submitted runs is zero: 22 in TREC 2003, 41 in TREC 2004, and 44 in TREC 2005.

An evaluation in which the median score for many questions is zero has many shortcomings. For one, it is difficult to tell if a particular run is "better" than another—even though they may be very different in other salient properties such as length, for example. The discriminative power of the present F-score measure is called into question: are present systems that bad, or is the current scoring model insufficient to discriminate between different (poorly performing) systems?

Also, as pointed out by Voorhees (2005), a score distribution heavily skewed towards zero makes meta-analysis of evaluation stability hard to perform. Since such studies depend on variability in scores, evaluations would appear more stable than they actually are.

While there are obviously shortcomings to the current scheme of labeling nuggets as either "vital" or "okay", the distinction does start to capture the intuition that "not all nuggets are created equal". Some nuggets are inherently more important than others, and this should be reflected in the evaluation methodology. The solution, we believe, is to solicit judgments from multiple assessors to develop a more refined sense of nugget importance. It is important, however, to balance the amount of additional manual effort required and the gains derived from those efforts. We present the idea of building "nugget pyramids", which addresses the shortcomings noted here, and then assess the implications of this new scoring model against data from TREC 2003, 2004, and 2005.

4 Building Nugget Pyramids

As previously pointed out (Lin and Demner-Fushman, 2005b), the question answering and multi-document summarization communities are converging on the task of addressing complex information needs from complementary perspectives. From an evaluation point of view, this provides opportunities for cross-fertilization and exchange of fresh ideas. As an example of intellectual intercourse, the recently-developed Pourpre metric for automatically evaluating answers to complex questions (Lin and Demner-Fushman, 2005a) employs *n*-gram overlap to compare system responses to reference output, an idea originally implemented in the Rouge metric for summarization evaluation (Lin and Hovy, 2003). Drawing additional inspiration from research on summarization evaluation, we adapt the pyramid evaluation scheme (Nenkova and Passonneau, 2004) to address the shortcomings of the vital/okay distinction in the nugget-based evaluation methodology.

The basic intuition behind the pyramid scheme (Nenkova and Passonneau, 2004) is simple: the importance of a fact is directly related to the number of people that recognize it as so (i.e., its popularity). The evaluation methodology calls for assessors to annotate Semantic Content Units (SCUs) found within model reference summaries. The weight assigned to an SCU is equal to the number of annotators that have marked the particular unit. These SCUs can be arranged in a pyramid, with the highest-scoring elements at the top: a "good" summary should contain SCUs from a higher level in the pyramid before a lower tier, since such elements are deemed "more vital".

	2003		**2004**		**2005**	
Assessor	Kendall's τ	zeros	Kendall's τ	zeros	Kendall's τ	zeros
0	1.00	22	1.00	41	1.00	44
1	0.908	20	0.933	36	0.888	43
2	0.896	21	0.916	43	0.900	41
3	0.903	21	0.917	38	0.897	39
4	0.912	20	0.914	42	0.879	56
5	0.873	23	0.926	40	0.841	53
6	0.889	29	0.908	32	0.894	39
7	0.900	22	0.930	37	0.890	54
8	0.909	18	0.932	29	0.891	35
9	0.879	26	0.908	49	0.877	58
average	0.896	22.2	0.920	38.7	0.884	46.2

Table 2: Kendall's τ correlation between system scores generated using "official" vital/okay judgments and each assessor's judgments. (Assessor 0 represents the original NIST assessors.)

This pyramid scheme can be easily adapted for question answering evaluation since a nugget is roughly comparable to a Semantic Content Unit. We propose to build nugget pyramids for answers to complex questions by soliciting vital/okay judgments from multiple assessors, i.e., take the original reference nuggets and asking different humans to classify each as either "vital" or "okay". The weight assigned to each nugget is simply equal to the number of different assessors that deemed the nugget vital. We then normalize the nugget weights (per-question) so that the maximum possible weight is one (by dividing each nugget weight by the maximum weight of that particular question). Therefore, a nugget assigned "vital" by the most assessors (not necessarily all) would receive a weight of one.[1]

The introduction of a more granular notion of nugget importance should be reflected in the calculation of F-score. We propose that nugget recall be modified to take into account nugget weight:

$$\mathcal{R} = \frac{\sum_{m \in A} w_m}{\sum_{n \in V} w_n}$$

Where A is the set of reference nuggets that are matched within a system's response and V is the set of all reference nuggets; w_m and w_n are the weights of nuggets m and n, respectively. Instead of a binary distinction based solely on matching vital nuggets, all nuggets now factor into the calculation of recall, subjected to a weight. Note that this new scoring model captures the existing binary vital/okay distinction in a straightforward way: vital nuggets get a score of one, and okay nuggets zero.

We propose to leave the calculation of nugget precision as is: a system would receive a length allowance of 100 non-whitespace characters for every nugget it retrieval (regardless of importance). Longer answers would be penalized for verbosity.

Having outlined our revisions to the standard nugget-based scoring method, we will proceed to describe our methodology for evaluating this new model and demonstrate how it overcomes many of the shortcomings of the existing paradigm.

5 Evaluation Methodology

We evaluate our methodology for building "nugget pyramids" using runs submitted to the TREC 2003, 2004, and 2005 question answering tracks (2003 "definition" questions, 2004 and 2005 "other"

[1] Since there may be multiple nuggets with the highest score, what we're building is actually a frustum sometimes. :)

questions). There were 50 questions in 2003 testset, 64 in 2004, and 75 in 2005. In total, there were 54 runs submitted to TREC 2003, 63 to TREC 2004, and 72 to TREC 2005. NIST assessors have manually annotated nuggets found in a given system's response, and this allows us to calculate the final F-score under different scoring models.

We recruited a total of nine different assessors for this study. Assessors consisted of graduate students in library and information science and computer science, as well as volunteers from the question answering community (obtained via a posting to NIST's TREC QA mailing list). Each assessor was given only the reference nuggets and asked to classify each as vital or okay. They were purposely asked to make these judgments without reference to documents in the corpus in order to expedite the assessment process—our goal is to propose a refinement to the current nugget evaluation methodology that addresses shortcomings while minimizing the amount of additional effort required. Combined with the answer key created by the original NIST assessors, we have a total of ten judgments for every single nugget in the three testsets.

We measure the correlation between system ranks generated by different scoring models using Kendall's τ, a commonly used rank correlation measure in information retrieval for quantifying the similarity between different scoring methods. Kendall's τ computes the "distance" between two rankings as the minimum number of pairwise adjacent swaps necessary to convert one ranking into the other. This value is normalized by the number of items being ranked such that two identical rankings produce a correlation of 1.0; the correlation between a ranking and its perfect inverse is -1.0; and the expected correlation of two rankings chosen at random is 0.0. Typically, a value of greater than 0.8 is considered "good", although 0.9 represents a threshold researchers generally aim for.

We hypothesize that system ranks are relatively unstable with respect to individual assessor's judgments. That is, how well a given system scores is to a large extent dependent on which assessor's judgments one uses for evaluation. This results from an inescapable fact of such evaluations, well known from studies of relevance in the information retrieval literature (Voorhees, 1998). Humans have legitimate differences in opinion regarding a nugget's importance, and there is no such thing as "the correct answer". However, we hypothesize that these variations can be smoothed out by building "nugget pyramids" in the manner we describe. Nugget weights reflect the combined judgments of many individual assessors, and scores generated with weights taken into account should correlate better with each individual assessor's opinion.

6 Results

To verify our hypothesis about the instability of using any individual assessor's judgments, we calculated the Kendall's τ correlation between system scores generated using the "official" vital/okay judgments (provide by NIST assessors) and each individual assessor's judgments. This is shown in Table 2. The original NIST judgments are listed as "assessor 0" (and not included in the averages). For all scoring models discussed in this paper, we set β, the parameter that controls the relative importance of precision and recall, to three.[2] Results show that although official rankings generally correlate well with rankings generated by our nine additional assessors, the agreement is far from perfect. Yet, in reality, the opinions of our nine assessors are not any less valid than those of the NIST assessors—NIST does not occupy a privileged position on what constitutes a good "definition". We can see that variations in human judgments do not appear to be adequately captured by the current scoring model.

Table 2 also shows the number of questions for which systems' median score was zero based on each individual assessor's judgments (out of 50 questions for TREC 2003, 64 for TREC 2004, and 75 for TREC 2005). These numbers are worrisome: in TREC 2004, for example, over half the questions (on average) have a median score of zero, and over three quarters of questions, according to assessor

[2] Note that $\beta = 5$ in the official TREC 2003 evaluation.

	2003	**2004**	**2005**
0	0.934	0.943	0.901
1	0.962	0.940	0.950
2	0.938	0.948	0.952
3	0.938	0.947	0.950
4	0.936	0.922	0.914
5	0.916	0.956	0.887
6	0.916	0.950	0.958
7	0.949	0.933	0.927
8	0.964	0.972	0.953
9	0.912	0.899	0.881
average	0.936	0.941	0.927

Table 3: Kendall's τ correlation between system rankings generated using the ten-assessor nugget pyramid and those generated using each individual assessor's judgments. (Assessor 0 represents the original NIST assessors.)

9. This is problematic for the various reasons discussed in Section 3.

To evaluate scoring models that combine the opinions of multiple assessors, we built "nugget pyramids" using all ten sets of judgments in the manner outlined in Section 4. All runs submitted to each of the TREC evaluations were then rescored using the modified F-score formula, which takes into account a finer-grained notion of nugget importance. Rankings generated by this model were then compared against those generated by each individual assessor's judgments. Results are shown in Table 3. As can be seen, the correlations observed are higher than those in Table 2, meaning that a nugget pyramid better captures the opinions of each individual assessor. A two-tailed t-test reveals that the differences in averages are statistically significant ($p << 0.01$ for TREC 2003/2005, $p < 0.05$ for TREC 2004).

What is the effect of combining judgments from different numbers of assessors? To answer this question, we built ten different nugget pyramids of varying "sizes", i.e., combining judgments from one through ten assessors. The Kendall's τ correlations between scores generated by each of these and scores generated by each individual assessor's judgments were computed. For each pyramid, we computed the average across all rank correlations, which captures the extent to which that particular pyramid represents the opinions of all ten assessors. These results are shown in Figure 2. The increase in Kendall's τ that comes from adding a second assessor is statistically significant, as revealed by a two-tailed t-test ($p << 0.01$ for TREC 2003/TREC 2005, $p < 0.05$ for TREC 2004), but ANOVA reveals no statistically significant difference beyond two assessors.

From these results, we can conclude that adding a second assessor yields a scoring model significantly better at capturing the variance in humans' relevance judgments. In this respect, little is gained beyond two assessors. If this is the only advantage provided by nugget pyramids, then the boost in rank correlations may not be sufficient to justify the extra manual effort. As we shall see, however, nugget pyramids offer other benefits as well.

Evaluation by our nugget pyramid greatly reduces the number of questions whose median score is zero. As previously discussed, a strict vital/okay split translates into a score of zero for systems that do not return any vital nuggets. However, nugget pyramids reflect a more refined sense of nugget importance, which results in fewer zero scores. Figure 3 shows the number of questions whose median score is zero (normalized as a fraction of the entire testset) by nugget pyramids built from a varying number of assessors. With four or more assessors, the number of questions whose median is zero for the TREC 2003 testset drops to 17; for TREC 2004, 23 for seven or more assessors; for TREC 2005, 27 for nine or more assessors. In other words, F-scores generated using our methodology are far more

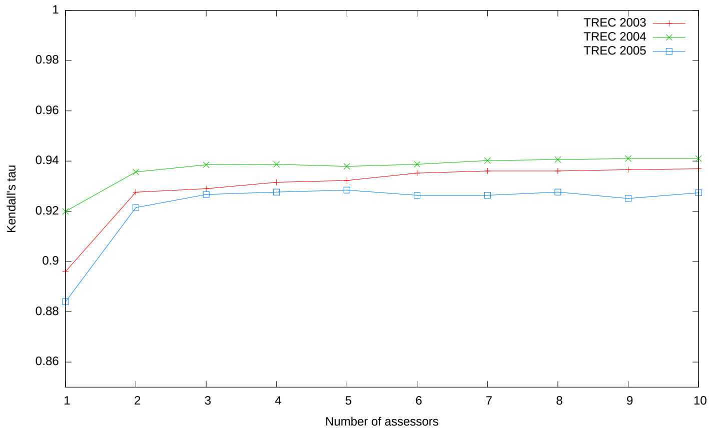

Figure 2: Average agreement (Kendall's τ) between individual assessors and nugget pyramids built from different numbers of assessors.

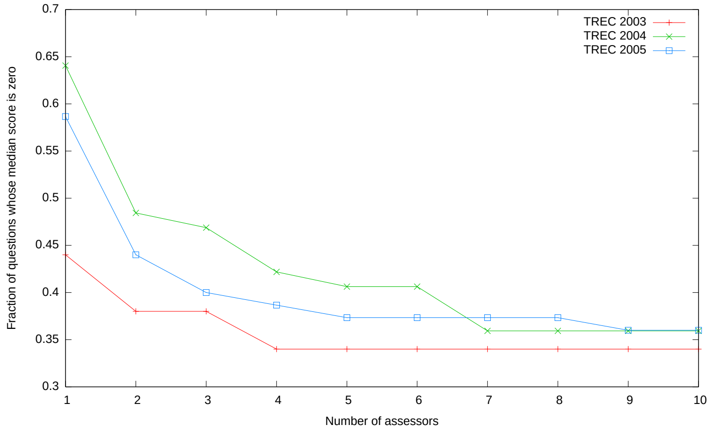

Figure 3: Fraction of questions whose median score is zero plotted against number of assessors whose judgments contributed to the nugget pyramid.

discriminative. The remaining questions with zero medians, we believe, reflect of the state of the art.

An example of a nugget pyramid that combines the opinions of all ten assessors is shown in Table 4 for the target "AARP". Judgments from the original NIST assessor are also shown (cf. Table 1). Note that there is a strong correlation between the original vital/okay judgments and the refined nugget weights based on the pyramid, indicating that (in this case, at least) the intuition of the NIST assessor matches that of the other assessors.

7 Discussion

In balancing the tradeoff between advantages provided by nugget pyramids and the additional manual effort necessary to create them, what is the optimal number of assessors to solicit judgments from? Results shown in Figures 2 and 3 provide some answers. In terms of better capturing different assessors' opinions, little appears to be gained from going beyond two assessors. However, adding more judgments does decrease the number of questions whose median score is zero, resulting in a more discriminative metric. Beyond five or some assessors, the number of questions with a zero median score remains relatively stable. We believe that around five assessors yield the smallest nugget pyramid that confers the advantages of the methodology.

1.0	vital	Largest seniors organization
0.9	vital	Membership eligibility is 50+
0.8	vital	30+ million members
0.7	vital	Largest dues paying organization
0.2	okay	Most of its work done by volunteers
0.1	okay	Spends heavily on research & education
0.1	okay	Receives millions for product endorsements
0.1	okay	Receives millions from product endorsements
0.0	okay	Abbreviated name to attract boomers

Table 4: Answer nuggets for the target "AARP"

The idea of building "nugget pyramids" in an extension of a similarly-named evaluation scheme in document summarization, although there are important differences. Nenkova and Passonneau (2004) call for multiple assessors to annotate SCUs, which is much more involved than the methodology presented here, where the nuggets are fixed and assessors only provide additional judgments about their importance. This obviously has the advantage of streamlining the assessment process, but has the potential to miss other important nuggets that were not identified in the first place. Our experimental results, however, suggest that this is a worthwhile tradeoff. The explicit goal of this work was to develop scoring models for nugget-based evaluation that would address shortcomings of the present approach, while introducing minimal overhead in terms of additional resource requirements. To this end, we have been successful.

Nevertheless, there are a number of issues that are worth mentioning. To speed up the assessment process, assessors were instructed to provide "snap judgments" given only the list of nuggets and the target. No additional context was provided, e.g., documents from the corpus or sample system responses. It is also important to note that the reference nuggets were never meant to be read by other people—NIST makes no claim for them to be well-formed descriptions of the facts themselves. These answer keys were primarily note-taking devices to assist in the assessment process. The important question, however, is whether scoring variations caused by poorly-phrased nuggets are smaller than the variations caused by legitimate inter-assessor disagreement regarding nugget importance. Our experiments appear to suggest that, overall, the nugget pyramid methodology is sound and can adequately cope with these difficulties.

8 Conclusion

The central importance that quantitative evaluation plays in advancing the state of the art in language technologies warrants close examination of evaluation methodologies themselves to ensure that they are measuring "the right thing". In this work, we have identified a shortcoming in the present nugget-based paradigm for assessing answers to complex questions. The vital/okay distinction was designed to capture the intuition that some nuggets are more important than others, but as we have shown, this comes at a cost in stability and discriminative power of the metric. We propose a revised model that incorporates judgments from multiple assessors in the form of a "nugget pyramid", and demonstrate how this addresses many of the previous shortcomings. Our model stands as the proposed metric for evaluating complex questions in the TREC 2006 question answering track.

9 Acknowledgments

This work has greatly benefited from discussions with Ellen Voorhees and participants at TREC 2005. We are grateful for the nine assessors who kindly provided nugget judgments that made this work possible. In addition, we would like to thank Kiri for her kind support.

References

Wesley Hildebrandt, Boris Katz, and Jimmy Lin. 2004. Answering definition questions with multiple knowledge sources. In *Proceedings of the 2004 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2004)*.

Jimmy Lin and Dina Demner-Fushman. 2005a. Automatically evaluating answers to definition questions. In *Proceedings of the 2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*.

Jimmy Lin and Dina Demner-Fushman. 2005b. Evaluating summaries and answers: Two sides of the same coin? In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2003)*.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the 2004 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2004)*.

Ellen M. Voorhees. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998)*.

Ellen M. Voorhees. 2003. Overview of the TREC 2003 question answering track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*.

Ellen M. Voorhees. 2005. Using question series to evaluate question answering system effectiveness. In *Proceedings of the 2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*.