LAMP-TR-119 February 2005
CS-TR-4695
UMIACS-TR-2005-04

Automatically Evaluating Answers to Definition Questions

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

Following recent developments in the automatic evaluation of machine translation and document summarization, we present a similar approach, implemented in a measure called POURPRE, for automatically evaluating answers to definition questions. Until now, the only way to assess the correctness of answers to such questions involves manual determination of whether an information nugget appears in a system's response. The lack of automatic methods for scoring system output is an impediment to progress in the field, which we address with this work. Experiments with the TREC 2003 and TREC 2004 QA tracks indicate that rankings produced by our metric correlate highly with official rankings, and that POURPRE outperforms a direct application of existing metrics.

Last updated: May 16, 2005
Keywords: question answering, definition questions, evaluation

1 Introduction

Recent interest in question answering has shifted away from factoid questions such as "What city is the home to the Rock and Roll Hall of Fame?", which can typically be answered by a short noun phrase, to more complex and difficult questions. One interesting class of information needs concerns so-called definition questions such as "Who is Vlad the Impaler?", whose answers would include "nuggets" of information about the 16th century warrior prince's life, accomplishments, and legacy. Actually a misnomer, definition questions can be better paraphrased as "Tell me interesting things about *X*.", where *X* can be a person, an organization, a common noun, etc. Taken another way, definition questions might be viewed as simultaneously asking a whole series of factoid questions about the same entity (e.g., "When was he born?", "What was his occupation?", "Where did he live?", etc.), except that these questions are not known in advance; see Prager et al. (2004) for an implementation based on this view of definition questions.

Much progress in natural language processing and information retrieval has been driven by the creation of reusable test collections. A test collection consists of a corpus, a series of well-defined tasks, and a set of judgments indicating the "correct answers". To complete the picture, there must exist meaningful metrics to evaluate progress, and ideally, a machine should be able to compute these values automatically. Although "answers" to definition questions are known, there is no way to automatically and objectively determine if they are present in a given system's response (we will discuss why in Section 2). The experimental cycle is thus tortuously long; to accurately assess the performance of new techniques, one must essentially wait for expensive, large-scale evaluations that employ human assessors to judge the runs. This situation mirrors the state of machine translation and document summarization research a few years ago. Since then, however, automatic scoring metrics such as BLEU and ROUGE have been introduced as stopgap measures to facilitate experimentation.

Following these recent developments in evaluation research, we propose POURPRE, a technique for automatically evaluating answers to definition questions. Like the abovementioned metrics, POURPRE is based on *n*-gram co-occurrences, but has been adapted for the unique characteristics of the question answering task. This paper will show that POURPRE can accurately assess the quality of answers to definition questions without human intervention, allowing experiments to be performed with rapid turnaround. We hope that this will enable faster exploration of the solution space and lead to accelerated advances in the state of the art.

This paper is organized as follows: In Section 2, we briefly describe how definition questions are evaluated, drawing attention to many of the intricacies involved. We discuss previous work in Section 3, relating POURPRE to evaluation metrics for other language applications. Section 4 discusses metrics for evaluating the quality of an automatic scoring algorithm. The POURPRE measure itself is outlined in Section 5; POURPRE scores are correlated with official human-generated scores in Section 6, and also compared to existing metrics. In Section 7, we explore the effect that judgment variability has on the stability of definition question evaluation, and its implications for automatic scoring algorithms.

2 Evaluating Definition Questions

To date, two formal evaluations of definition questions have been conducted, at TREC 2003 and TREC 2004.[1] In this section, we describe the setup of the task and the evaluation methodology.

Answers to definition questions are comprised of an unordered set of [document-id, answer string] pairs, where the strings are presumed to provide some relevant information about the entity being "defined", usually called the target. Although no explicit limit is placed on the length of the answer string, the final scoring metric penalizes verbosity (discussed below).

[1] TREC 2004 questions were arranged around "topics"; definition questions were implicit in the "other" questions.

1	*vital*	32 kilograms plutonium powered
2	*vital*	seven year journey
3	*vital*	Titan 4-B Rocket
4	*vital*	send Huygens to probe atmosphere of Titan, Saturn's largest moon
5	*okay*	parachute instruments to planet's surface
6	*okay*	oceans of ethane or other hydrocarbons, frozen methane or water
7	*vital*	carries 12 packages scientific instruments and a probe
8	*okay*	NASA primary responsible for Cassini orbiter
9	*vital*	explore remote planet and its rings and moons, Saturn
10	*okay*	European Space Agency ESA responsible for Huygens probe
11	*okay*	controversy, protest, launch failure, re-entry, lethal risk, humans, plutonium
12	*okay*	Radioisotope Thermoelectric Generators, RTG
13	*vital*	Cassini, NASA'S Biggest and most complex interplanetary probe
14	*okay*	find information on solar system formation
15	*okay*	Cassini Joint Project between NASA, ESA, and ASI (Italian Space Agency)
16	*vital*	four year study mission

Table 1: The "answer key" to the question "What is the Cassini space probe?"

[XIE19971012.0112] The Cassini space probe, due to be launched from Cape Canaveral in Florida of the United States tomorrow, has a 32 kilogram plutonium fuel payload to power its seven year journey to Venus and Saturn.
Nuggets assigned: 1, 2

[NYT19990816.0266] Early in the Saturn visit, Cassini is to send a probe named Huygens into the smog-shrouded atmosphere of Titan, the planet's largest moon, and parachute instruments to its hidden surface to see if it holds oceans of ethane or other hydrocarbons over frozen layers of methane or water.
Nuggets assigned: 4, 5, 6

Figure 1: Examples of judging actual system responses.

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

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

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.

Figure 2: Official definition of F-measure.

To evaluate system responses, NIST pools answer strings from all systems, 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"—a list of "information nuggets" about the target. An information 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). The assessor then manually classifies all nuggets 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.[2] As an example, nuggets for the question "What is the Cassini space probe?" are shown in Table 1.

Once this answer key of vital/okay nuggets is created, the assessor then manually scores each run. For each system response, he or she decides whether or not each nugget is present. Assessors do not simply perform string matches in this decision process; rather, this matching occurs at the conceptual level, abstracting away from issues such as vocabulary differences, syntactic divergences, paraphrases, etc. Two examples of this matching process are shown in Figure 1: nuggets 1 and 2 were found to be in the top passage, while nuggets 4, 5, and 6 were found to be in the bottom passage. It is exactly this process of conceptually matching nuggets from the answer key with system responses that we are attempting to capture with an automatic scoring algorithm.

The final F-score for an answer is calculated in the manner described in Figure 2, and the final score of a run is simply the average across the scores of all questions. The metric 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). Nugget recall is calculated solely on vital nuggets, while nugget precision is approximated by a length allowance given based on the number of both vital and okay nuggets returned. Early on in a pilot study, researchers discovered that it was impossible for assessors to consistently enumerate the total set of nuggets contained in a system response, given that they are usually extracted text segments from documents (Voorhees, 2003). Thus, a penalty for verbosity serves as a surrogate for precision.

[2]For a critique of this distinction, please refer to (Hildebrandt et al., 2004).

3 Previous Work

The idea of employing *n*-gram co-occurrence statistics to score the output of a computer system against one or more desired reference outputs was first successfully implemented in the BLEU metric for machine translation (Papineni et al., 2002). Since then, the basic method for scoring translation quality has been improved upon by others, e.g., (Babych and Hartley, 2004; Lin and Och, 2004a; Lin and Och, 2004b). The basic idea has been extended to evaluating document summarization with ROUGE (Lin and Hovy, 2003).

Recently, Soricut and Brill (2004) employed *n*-gram co-occurrences to evaluate question answering in a FAQ domain; unfortunately, the task differs from definition question answering, making their results not directly applicable. Xu et al. (2004) applied ROUGE to automatically evaluate answers to definition questions, viewing the task as a variation of document summarization. Because TREC answer nuggets were terse phrases, the authors found it necessary to rephrase them—two humans were asked to manually create "reference answers" based on the assessors' nuggets and IR results, which was a labor-intensive process. Furthermore, Xu et al. did not perform a large-scale assessment of the reliability of ROUGE for evaluating definition answers.

4 Criteria for Success

Before proceeding to our description of POURPRE, it is important to first define the basis for evaluating the quality of an automatic evaluation algorithm. Correlation between official scores and automatically-generated scores, as measured by the coefficient of determination R^2, seems like an obvious metric for quantifying the performance of a scoring algorithm. Indeed, this measure has been employed in the evaluation of BLEU, ROUGE, and other related metrics.

However, we believe that there are better measures of performance. In comparative evaluations, we ultimately want to determine if one technique is "better" than another (recognizing that absolute scores are often relatively meaningless). Thus, the system rankings produced by a particular scoring method are often more important than the actual scores themselves. Following the information retrieval literature, we employ Kendall's τ to capture this insight. 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; cf. (Voorhees and Tice, 1999). Typically, a value of greater than 0.8 is considered "good", although 0.9 represents a threshold researchers generally aim for. In this study, we primarily focus on Kendall's τ, but also report R^2 values where appropriate.

5 Pourpre

Previously, it has been assumed that matching nuggets from the assessors' answer key with systems' responses must be performed manually because it involves semantics (Voorhees, 2003). We would like to challenge this assumption and hypothesize that term co-occurrence statistics can serve as a surrogate for this semantic matching process. Experience with the ROUGE metric has demonstrated the effectiveness of matching unigrams, an idea we employ in our POURPRE metric. We hypothesize that matching bigrams, trigrams, or any other longer *n*-grams will not be beneficial, because they primarily account for the fluency of a response, more relevant in a machine translation task. Since answers to definition questions are usually document extracts, fluency is less important a concern. This hypothesis will be verified in Section 6.1.

The idea behind POURPRE is relatively straightforward: match nuggets by summing the unigram co-occurrences between terms from each nugget and terms from the system response. We decided to start with the simplest possible approach: count the word overlap and divide by the total number of terms in the answer nugget. The only additional wrinkle is to ensure that all words appear within the same answer string. Since nuggets represent coherent concepts, they are unlikely to be spread across different answer strings (which are usually different extracts of source documents). As a simple example, let's say we're trying to determine if the nugget "A B C D" is contained in the following system response:

1. A
2. B C D
3. D
4. A D

The match score assigned to this nugget would be 3/4, from answer string 2; no other answer string would get credit for this nugget. This provision reduces the impact of coincidental term matches.

Once we determine the match score for every nugget, the final F-score is calculated in the usual way, except that the automatically-calculated match scores are substituted where appropriate. For example, nugget recall now becomes the sum of the match scores for all vital nuggets divided by the total number of vital nuggets. In the official F-score calculation, the length allowance—for the purposes of computing nugget precision—was 100 non-whitespace characters for every okay and vital nugget returned. Since nugget match scores are now fractional, this required some adjustment. We settled on an allowance of 100 non-whitespace characters for every nugget match that had non-zero score.

A major drawback of this basic unigram overlap approach is that all terms are considered equally important—surely, matching "year" in a system's response should count for less than matching "Huygens", in the example about the Cassini space probe. We decided to capture this intuition using inverse document frequency, a commonly-used measure in information retrieval; $idf(t_i)$ is defined as $log(N/c_i)$, where N is the number of documents in the collection, and c_i is the number of documents that contain the term t_i. With scoring based on *idf*, term counts are simply replaced with *idf* sums in computing the match score, i.e., the match score of a particular nugget is the sum of the *idf*s of matching terms in the system response divided by the sum of all term *idf*s from the answer nugget. To lessen the impact of spurious nugget matches based on coincidental co-occurrence of low *idf* terms, all normalized match scores lower than 0.005 were treated as zero.

Finally, we examined the effects of stemming on score quality. We compared matching stemmed terms from the nuggets and system responses, derived from the Porter stemmer (Porter, 1980), with matching the original terms.

In the next section, experiments with submissions from TREC 2003 and TREC 2004 are reported. We attempted two different methods for aggregating results: microaveraging and macroaveraging. For microaveraging, scores were calculated by computing the nugget match scores over all nuggets for all questions. For macroaveraging, scores for each question were first computed, and then averaged across all questions in the testset. With microaveraging, each nugget is given equal weight, while with macroaveraging, each question is given equal weight.

As a baseline, we revisited experiments by Xu et al. (2004) in using ROUGE to evaluate definition questions. What if we simply concatenated all the answer nuggets together and used the result as the "reference summary" (instead of using humans to create custom reference answers)? As an additional baseline, we compared POURPRE against the BLEU/NIST metric used in machine translation evaluation.

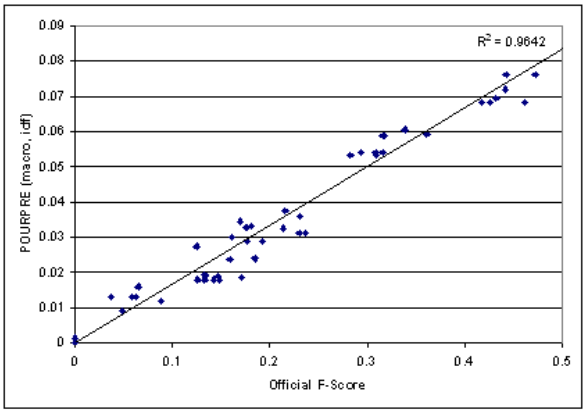

Figure 3: Scatter graph of official scores plotted against the POURPRE scores (macroaveraging, *idf* term weighting) for TREC 2003 ($\beta = 5$).

6 Evaluation of Pourpre

We evaluated all definition question runs submitted to the TREC 2003[3] and TREC 2004 question answering tracks with different variants of our POURPRE metric, and then compared the results with the official F-scores generated by human assessors. The Kendall's τ correlations between rankings produced by POURPRE and the official rankings are shown in Table 2. The coefficients of determination (R^2) between the two sets of scores are shown in Table 3. We report four separate variants along two different parameters: scoring by term counts only vs. scoring by term *idf*, and microaveraging vs. macroaveraging. A scatter graph plotting official F-scores against POURPRE scores (macroaveraging, *idf* term weighting) for TREC 2003 ($\beta = 5$) is shown in Figure 3. Corresponding graphs for other variants appear similar, and are not shown here.

The effect of stemming on the Kendall's τ correlation between POURPRE (macroaveraging, *idf* term matching) and official scores in shown in Table 4. For TREC 2004 ($\beta = 3$), stemming was found to decrease the correlation, but for TREC 2003 ($\beta = 3$ and $\beta = 5$), stemming appears to improve performance. Results from the same stemming experiment on the other POURPRE variants are similarly inconclusive.

For TREC 2003 ($\beta = 5$), we performed an analysis of rank swaps between official and POURPRE scores. A rank swap is said to have occurred if the relative ranking of two runs is different under different conditions—they are significant because rank swaps might prevent researchers from confidently drawing conclusions about the relative effectiveness of different techniques. We observed 83 rank swaps (out of a total of 1431 pairwise comparisons for 54 runs). A histogram of these rank swaps, binned by the difference in official score, is shown in Figure 4. As can be seen, 43 rank swaps (51.8%) occurred when the difference in official score is less than 0.02; there were no rank swaps observed for runs in which the official scores differed by more than 0.067. Since measurement error is an inescapable fact of evaluation, we need not be concerned with rank swaps that can be attributed to this factor. For TREC 2003, Voorhees (2003) calculated this value to be approximately 0.1; that is, in order to conclude with 95% confidence that one run is better than another, an absolute F-score difference greater than 0.1 must be observed. As can be seen, all the rank swaps observed can be attributed to error inherent in the evaluation process.

From these results, we can see that evaluation of definition questions is relatively coarse-grained. However, TREC 2003 was the first formal evaluation of definition questions; as methodologies are

[3]In TREC 2003, the value of β was arbitrarily set to five, which was later determined to favor recall too heavily. As a result, it was readjusted to three in TREC 2004. In our experiments with TREC 2003, we report figures for both values.

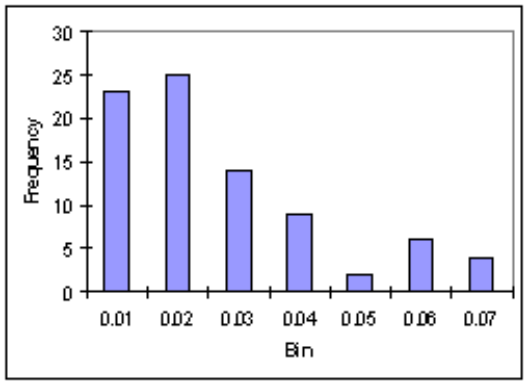

Figure 4: Histogram of rank swaps for TREC 2003 ($\beta = 5$), binned by difference in official score.

refined, the margin of error should go down. Although a similar error analysis for TREC 2004 has not been performed, we expect that most rank swaps between POURPRE and the official scores can still be attributed to measurement error inherent in evaluations.

Given the simplicity of our POURPRE metric, the correlation between our automatically-derived scores and the official scores is remarkable. Starting from a set of questions and a list of relevant nuggets, POURPRE can accurately assess the performance of a definition question answering system without any human intervention.

6.1 Comparison Against Bleu and Rouge

Since POURPRE uses many of the same ideas captured in BLEU and ROUGE, it would be worthwhile to use these existing metrics as baselines for comparison. Conceptually, however, the task of answering definition questions is closer to summarization then it is to machine translation, in that both are recall-oriented. Since the majority of question answering systems employ extractive techniques, fluency (i.e., precision) is not usually an issue in the evaluation.

Table 5 shows the Kendall's τ correlations between official rankings and rankings produced by the BLEU/NIST metrics[4] when directly applied to definition question answering. The best performance was achieved by unigrams, and far underperforms our POURPRE metric.

How does POURPRE stack up against using ROUGE[5] to directly evaluate definition questions? The Kendall's τ correlations between rankings produced by ROUGE (with and without stopword removal) and the official rankings are shown in Table 6. In all cases, ROUGE does not perform as well. Furthermore, ROUGE has an additional downside in that the metric cannot inform system developers *why* an answer received a particular score (in terms of which nuggets were found). Contrary to the work of Xu et al. (2004), this experiment shows that answer nuggets can be directly used as a "reference summary" in scoring the answers to definition questions; the manual creation of more coherent reference answers does not appear to be necessary.

We believe that POURPRE better correlates with official scores because it takes into account special characteristics of the task: the distinction between vital and okay nuggets, the length penalty, etc. Other than a higher correlation, POURPRE offers an advantage over ROUGE in that it provides a better diagnostic than a coarse-grained score. With our measure, it is possible to reconstruct which answer nuggets were present in a response and which nuggets were omitted; this allows researchers to conduct failure analyses to identify opportunities for improvement.

[4]We used version 11 of the evaluation software.

[5]We used ROUGE-1.4.2 with n set to 1, i.e. unigram matching, and maximum matching score rating.

Run	micro, count	macro, count	micro, *idf*	macro, *idf*
TREC 2004 ($\beta = 3$)	0.785	0.833	0.806	0.813
TREC 2003 ($\beta = 3$)	0.846	0.886	0.848	0.876
TREC 2003 ($\beta = 5$)	0.889	0.878	0.859	0.875

Table 2: Correlation (Kendall's τ) between rankings generated by POURPRE and official scores.

Run	micro, count	macro, count	micro, *idf*	macro, *idf*
TREC 2004 ($\beta = 3$)	0.837	0.929	0.904	0.914
TREC 2003 ($\beta = 3$)	0.919	0.963	0.941	0.957
TREC 2003 ($\beta = 5$)	0.954	0.965	0.957	0.964

Table 3: Correlation (R^2) between values generated by POURPRE and official scores.

Run	unstemmed	stemmed
TREC 2004 ($\beta = 3$)	0.813	0.795
TREC 2003 ($\beta = 3$)	0.876	0.875
TREC 2003 ($\beta = 5$)	0.875	0.871

Table 4: The effect of stemming on Kendall's τ; all runs with (macro, *idf*) variant of POURPRE.

Run	BLEU	NIST
TREC 2004 ($\beta = 3$)	0.168	0.235
TREC 2003 ($\beta = 3$)	0.169	0.207
TREC 2003 ($\beta = 5$)	0.154	0.192

Table 5: Correlation (Kendall's τ) between rankings generated by BLEU/NIST and official scores.

Run	+stop	−stop
TREC 2004 ($\beta = 3$)	0.780	0.786
TREC 2003 ($\beta = 3$)	0.780	0.816
TREC 2003 ($\beta = 5$)	0.807	0.843

Table 6: Correlation (Kendall's τ) between rankings generated by ROUGE and official scores.

Run	everything vital	vital/okay flipped	random judgments
TREC 2004 ($\beta = 3$)	0.831	0.765	0.797 $\pm$ 0.054
TREC 2003 ($\beta = 3$)	0.883	0.804	0.854 $\pm$ 0.023
TREC 2003 ($\beta = 5$)	0.904	0.824	0.868 $\pm$ 0.021

Table 7: Correlation (Kendall's τ) between scores under different variations of judgments and the official scores. The 95% confidence interval is presented for the random judgments case.

7 The Effect of Variability in Judgments

As with many other information retrieval tasks, legitimate differences in opinion about relevance are an inescapable fact of evaluating definition questions—systems are designed to satisfy real-world information needs, and users inevitably disagree on what nuggets are important or relevant. These disagreements manifest as scoring variations in an evaluation setting. The important issue, however, is the degree to which variations in judgments affect conclusions that can be drawn in a comparative evaluation, i.e., can we still confidently conclude that one system is "better" than another? For the *ad hoc* document retrieval task, research has shown that system rankings are stable with respect to disagreements about document relevance (Voorhees, 2000). In this section, we explore the effect of judgment variability on the stability and reliability of TREC definition question answering evaluations.

The vital/okay distinction on nuggets is one major source of differences in opinion, as has been pointed out previously (Hildebrandt et al., 2004). In the Cassini space probe example, we disagree with the assessors' assignment in many cases. More importantly, however, there does not appear to be any operationalizable rules for classifying nuggets as either vital or okay. Without any guiding principles, how can we expect our systems to automatically recognize this distinction, which has significant impact on scoring?

How do differences in opinion about vital/okay nuggets impact the stability of system rankings? To answer this question, we measured the Kendall's τ correlation between the official rankings and rankings produced by different variations of the answer key. Three separate variants were considered:

- all nuggets considered vital
- vital/okay flipped (all vital nuggets become okay, and all okay nuggets become vital)
- randomly assigned vital/okay labels

Results are shown in Table 7. For the last condition, we conducted one thousand random trials, taking into consideration the original distribution of the vital and okay nuggets for each question using a simplified version of the Metropolis-Hastings algorithm (Chib and Greenberg, 1995); the 95% confidence intervals for this experiment are reported. We only show results with macroaveraging and *idf* term weighting; values for other POURPRE variants appear similar.

These results suggest that system rankings are sensitive to assessors' difference in opinion about what constitutes a vital or okay nugget. In general, the Kendall's τ values observed here are lower than values computed from corresponding experiments in *ad hoc* document retrieval (Voorhees, 2000). To illustrate, the distribution of ranks for the top two runs from TREC 2004 (RUN-12 and RUN-8) over the one thousand random trials is shown in Figure 5. In 814 trials, RUN-12 was ranked as the highest-scoring run; however, in 152 trials, RUN-8 was ranked as the highest-scoring run, a small but significant figure. Factoring in differences of opinion about the vital/okay distinction, one could not conclude with certainty which was the "best" run in the evaluation.

It appears that differences between POURPRE and the official scores are about the same as (or in some cases, smaller than) differences between the official scores and scores based on variant answer

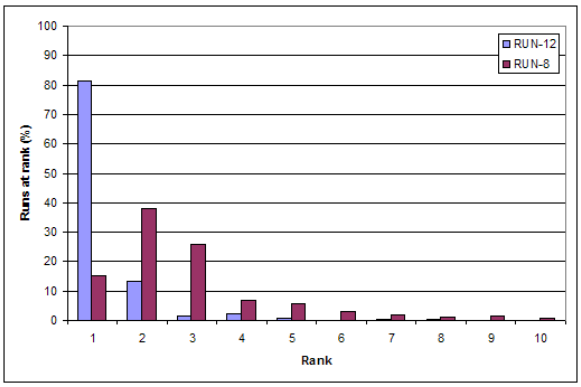

Figure 5: Distribution of rank placement using random judgments (for top two runs from TREC 2004).

keys. This means that further refinement of the metric to increase correlation with human-generated scores would not be particularly meaningful; it would essentially be overtraining on the whims of a particular human assessor. We believe that sources of judgment variability and techniques for managing it represent important areas for future study.

8 Conclusion

We hope that POURPRE can accomplish for definition question answering what BLEU has done for machine translation, and ROUGE for document summarization: allow laboratory experiments to be conducted with rapid turnaround. A much shorter experimental cycle will allow researchers to explore different techniques and receive immediate feedback on their effectiveness. Hopefully, this will translate into rapid progress in the state of the art.[6]

9 Acknowledgements

We would like to thank Donna Harman and Bonnie Dorr for comments on earlier drafts of this paper. The first author would like to thank Kiri for her kind support.

References

Bogdan Babych and Anthony Hartley. 2004. Extending the BLEU MT evaluation method with frequency weightings. In *Proceedings of ACL 2004*.

Siddhartha Chib and Edward Greenberg. 1995. Understanding the Metropolis-Hastings algorithm. *American Statistician*, 49(4):329–345.

Wesley Hildebrandt, Boris Katz, and Jimmy Lin. 2004. Answering definition questions with multiple knowledge sources. In *Proceedings of HLT/NAACL 2004*.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT/NAACL 2003*.

[6] A toolkit implementing the POURPRE metric can be downloaded at http://www.umiacs.umd.edu/~jimmylin/downloads/

Chin-Yew Lin and Franz Josef Och. 2004a. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of ACL 2004*.

Chin-Yew Lin and Franz Josef Och. 2004b. ORANGE: A method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of COLING 2004*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*.

Martin Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

John Prager, Jennifer Chu-Carroll, and Krzysztof Czuba. 2004. Question answering using constraint satisfaction: QA–by–Dossier–with–Constraints. In *Proceedings of ACL 2004*.

Radu Soricut and Eric Brill. 2004. A unified framework for automatic evaluation using n-gram co-occurrence statistics. In *Proceedings of ACL 2004*.

Ellen M. Voorhees and Dawn M. Tice. 1999. The TREC-8 question answering track evaluation. In *Proceedings of TREC-8*.

Ellen M. Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36(5):697–716.

Ellen M. Voorhees. 2003. Overview of the TREC 2003 question answering track. In *Proceedings of TREC 2003*.

Jinxi Xu, Ralph Weischedel, and Ana Licuanan. 2004. Evaluation of an extraction-based approach to answering definition questions. In *Proceedings of SIGIR 2004*.