Methods for Automatically Evaluating Answers to Complex Questions*

Jimmy Lin[1,2,3] and Dina Demner-Fushman[2,3]
[1]College of Information Studies
[2]Department of Computer Science
[3]Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742, USA
jimmylin@umd.edu, demner@cs.umd.edu

Abstract

Evaluation is a major driving force in advancing the state of the art in language technologies. In particular, methods for automatically assessing the quality of machine output is the preferred method for measuring progress, provided that these metrics have been validated against human judgments. Following recent developments in the automatic evaluation of machine translation and document summarization, we present a similar approach, implemented in a measure called POURPRE, an automatic technique for evaluating answers to complex questions based on n-gram co-occurrences between machine output and a human-generated answer key. 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, TREC 2004, and TREC 2005 QA tracks indicate that rankings produced by our metric correlate highly with official rankings, and that POURPRE outperforms direct application of existing metrics.

1 Introduction

Question answering, which lies at the intersection between natural language processing and information retrieval, has recently emerged as a technology that promises to deliver "answers" instead of "hits". In the past few years, researchers have made significant strides in systems capable of answering fact-based natural language questions such as "What city is the home to the Rock and Roll Hall of Fame?", "Who invented the paper clip?", and "How far is it from the pitcher's mound to home plate?" These so-called "factoid" questions can be typically answered with named entities such as dates, locations, people, organizations, measures, etc.

Implicit in the factoid question answering task is the notion that there exists a single, short, correct answer. A straightforward extension that relaxes this criterion yields list questions such as "What countries export oil?", where the desired response is an unordered set of named entities. List questions are challenging because it is difficult to predict *a priori* the total number of existing answer instances and to recognize answers that refer to the same entity (due to synonymy, name variations, etc.).

*Please cite as: Jimmy Lin and Dina Demner-Fushman. Methods for Automatically Evaluating Answers to Complex Questions. *Information Retrieval*, 9(5):565–587, 2006. This is the pre-print version of a published article. Citations to and quotations from this work should reference that publication. If you cite this work, please check that the published form contains precisely the material to which you intend to refer. (submitted: November 19, 2005. accepted: May 1, 2006)

Progress in factoid question answering does not alter the fact that such questions comprise only a small fraction of all possible information needs. To address the limitations of current technology, researchers have begun to explore techniques for answering more complex questions such as the following:

- Who is Aaron Copland?
- How have South American drug cartels been using banks in Liechtenstein to launder money?
- What was the Pentagon panel's position with respect to the dispute over the US Navy training range on the island of Vieques?

The first is an example of a so-called "definition" question, where the goal is to generate a profile of a person, entity, or event that integrates information from multiple sources within a given text collection. The second is an example of a "relationship" question, focused on the ties (economic, familial, etc.) between different entities. The last is an example of a so-called "opinion" question, which might involve sentiment detection and analysis of language use. This work presents a method for automatically evaluating answers to such complex questions.

Research in evaluation methodology is important because evaluation is arguably the single biggest force that drives advances in the state of the art. Typically, quantitative evaluation of language technology is accomplished through the use of a test collection, which consists of a corpus, a series of well-defined tasks, and a set of judgments indicating the desired system output. Another requirement is the existence of a meaningful metric that quantifies machine performance in relation to a reference, usually human output.

Automatic methods for evaluating system output are highly desirable because humans represent a bottleneck in the experimental cycle. The ability to assess system performance without human intervention allows quick experimental turnaround and rapid exploration of the solution space, which often leads to dramatic advances in the state of the art. A case in point is machine translation, where the development of BLEU (Papineni et al., 2002) and related metrics has stimulated significant improvements in translation quality. Prior to the existence of these automatic metrics, humans were required to assess translation quality, a process that was both slow and error-prone.

This work presents POURPRE, an automatic technique for evaluating answers to complex questions based on n-gram co-occurrences between machine output and a human-generated answer key. Experiments with data from the TREC question answering evaluations show that our metric correlates well with human judgments. POURPRE represents the first automatic evaluation method for complex questions: by serving as a surrogate for human assessors, it enables rapid experimentation in question answering, which will hopefully lead to accelerated advances in the state of the art.

This paper is organized as follows: Section 2 discusses the general methodology for evaluating answers to complex questions, and how it is implemented in the context of the TREC question answering tracks. Section 3 relates our proposed metric with previous work for automatically evaluating language technologies: machine translation and document summarization. Section 4 provides an algorithmic description of POURPRE. Section 5 outlines our evaluation methodology for validating the metric using data from previous TREC question answering tracks. Section 6 presents results from our evaluation, comparing different variants of POURPRE and direct application of existing metrics. Section 7 describes a series of more detailed experiments that attempts to provide a better understanding of how POURPRE works and factors that affect its performance. Finally, Section 8 concludes this paper.

2 Evaluating Answers to Complex Questions

Consider a "definition" question such as "Who is Vlad the Impaler?" A "good" answer might include information about the 16th century warrior prince's life, accomplishments, and legacy. Actually a

misnomer, such questions can be better paraphrased as "Tell me interesting things about X.", where X (the target) can be a person, an organization, a common noun, an event, etc. Given that named entities alone are insufficient to describe the target, answers to "definition" questions might be comprised of natural language sentences and phrases that express relevant facts about the target entity. Taken together, these facts might qualify for a "good answer". Consider the following "nuggets" of information about "Vlad the Impaler":

16th century warrior prince
Inspiration for Bram Stoker 1897 novel "Dracula"
Buried in medieval monastery on islet in Lake Snagov
Impaled opponents on stakes when they crossed him
Lived in Transylvania (Romania)
Fought Turks
Called "Prince of Darkness"
Possibly related to British royalty

By viewing complex questions as requests for collections of facts that address a particular information need, it becomes possible to characterize the properties of a "good" answer. From this basic idea, one can develop the desiderata for an evaluation metric:

- Answers that contain more relevant facts are preferred over answers that contain fewer relevant facts.
- Shorter answers are preferred over longer answers containing the same number of facts. Verbosity should be punished.
- Some facts are more important than others, and this should be reflected accordingly.

Despite differences in form, a fact-based evaluation methodology can be applied to many types of complex questions, including the "relationship" and "opinion" questions described above. In fact, this methodology has been implemented in the TREC question answering tracks, organized by the National Institute of Standards and Technology, since 2003. The QA tracks, which first began in 1999, provide the infrastructure and support necessary to conduct large-scale evaluations on shared collections using common testsets. This setup provides meaningful comparisons between different retrieval techniques. TREC evaluations have gained widespread acceptance as the *de facto* benchmark by which question answering performance is measured, and results from previous years provide data necessary to validate our automatic evaluation metric.

To date, NIST has conducted four separate evaluations of complex questions, all using a fact-based evaluation methodology: "definition/other" questions in TREC 2003–2005 (e.g., "What is feng shui?"); "relationship" questions in TREC 2005 (e.g., "Do the military personnel exchanges between Israel and India show an increase in cooperation? If so, what are the driving factors behind this increase?"); a small-scale pilot of "opinion" questions in 2005 (e.g., "Do the American people think that Elian Gonzalez should be returned to his father?"). Since relatively few teams participated in the TREC 2005 evaluation of relationship questions, this work focuses only on the three years' worth of definition/other questions.

In 2004 and 2005, factoid and list questions were organized in "series" around topics (Voorhees, 2005), for example:

the band Nirvana

1	factoid	Who is the lead singer/musician in Nirvana?
2	list	Who are the band members?

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?"

3	factoid	When was the band formed?
4	factoid	What is their biggest hit?
5	list	What are their albums?
6	factoid	What style of music do they play?
7	other	

Requests for definitions were implicit in the other questions associated with each topic. These other questions differed from the TREC 2003 definition questions only in that duplicate answers to factoid or list questions in the same series were not considered relevant; i.e., these questions are best paraphrased as "Tell me interesting things about X that I haven't already explicitly asked about." For the purposes of this study, this fine distinction was ignored, as the decision does not have any significant consequences for our results.

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", i.e., the target. Although no explicit upper limit is placed on the length of the answer string or the number of response pairs, the final F-score penalizes verbosity (described below).

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" 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). 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 for scoring, demonstrated below. As an example, nuggets for the question "What is the Cassini space probe?" are shown in Table 1.

[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 nuggets found in a system response, as determined by a human assessor.

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 2: Official definition of F-score.

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. Two examples of this matching process are shown in Figure 1: nuggets 1 and 2 were found in the top passage, while nuggets 4, 5, and 6 were found in the bottom passage.

The final F-score for an answer is calculated in the manner described in Figure 2, and the final score of a system run is simply 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 nuggets (which means no credit is given for returning okay nuggets), while nugget precision is approximated by a length allowance 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 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.

Because the answer key to a question only needs to be created once, the process of manually assessing each system response is the bottleneck in TREC evaluations. It is thought that a human is

required to determine the presence or absence of a particular nugget because assessors do not simply perform string matches in this decision process. Rather, matching occurs at the conceptual level, abstracting away from issues such as vocabulary differences, syntactic divergences, paraphrases, etc. Consider the following examples:

Who is Al Sharpton?
Nugget: Harlem civil rights leader
System response: New York civil rights activist

What is Freddie Mac?
Nugget: Fannie Mae is sibling of Freddie Mac
System response: both Fannie Mae and Freddie Mac, the quasi-governmental agencies that together ...

Who is Ari Fleischer?
Nugget: Elizabeth Dole's Press Secretary
System response: Ari Fleischer, spokesman for ... Elizabeth Dole

What is the medical condition shingles?
Nugget: tropical [*sic*] capsaicin relieves pain of shingles
System response: Epilepsy drug relieves pain from ... shingles

In each of the above cases, is the nugget contained in the system response fragment? Regardless of the judgment, these examples show the supposed necessity of having a human in the loop—to decide whether two strings contain the same semantic content. For this reason, there exists no automatic method for evaluating answers to complex questions. As a result, the experimental cycle for developing systems has been tortuously long; to accurately assess the performance of new techniques, one must essentially wait for the yearly TREC cycle (or conduct less-accurate in-house manual evaluations).

In attempts to remedy this situation, we reexamined the assumption that matching answer nuggets with system responses requires human involvement. We show that, despite inherent difficulties in matching meaning, automatically computing substring overlap is a viable alternative to human–in–the–loop evaluation. This insight, first demonstrated by Papineni et al. (2002) in the context of machine translation, has given rise to much subsequent work focused on the development of automatic evaluation metrics for various language technologies. POURPRE, our algorithm for evaluating answers to complex questions, follows very much in this line of research. It is shown through experiments on data from previous TRECs that POURPRE scores correlate well with human judgments, thus validating the usefulness of our metric for guiding system development.

3 Related Work

The evaluation of many language applications involves comparing system-generated output with one or more desired outputs, usually generated by humans. In machine translation, for example, testsets include a number of human-generated "reference" translations; in document summarization, manually-generated summaries provide examples of "good summaries". The idea of employing n-gram co-occurrence statistics (i.e., substring matches) to compare machine-generated output with desired output was first successfully implemented in the BLEU metric (Papineni et al., 2002) and the closely-related NIST metric (Doddington, 2002) for machine translation evaluation. Experiments have shown that these automatically-generated scores correlate well with human judgments across large testsets (although not on a per-sentence basis) (Kulesza and Shieber, 2004). Since then, the basic method for scoring translation quality has been improved upon by others, e.g., (Babych and Hartley, 2004; Lin and Och, 2004). Although there have been recent attempts to introduce more linguistically-rich features in

the evaluation algorithms, e.g., (Banerjee and Lavie, 2005), the basic idea of matching n-grams remains at the core of machine translation evaluation.

The simple idea of matching substrings from system output against a reference standard has also been applied to document summarization evaluation. The ROUGE metric (Lin and Hovy, 2003) has been shown to correlate well with human judgments and has been widely adopted by researchers for system development. Although more sophisticated methods, e.g., the pyramid scheme (Nenkova and Passonneau, 2004), have been proposed, ROUGE remains an important benchmark due to its ease of use and accuracy.

Recently, Soricut and Brill (2004) employed n-gram co-occurrences to evaluate question answering in a FAQ domain; unfortunately, the task differs from that of answering "definition" questions, making their results not directly comparable. 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 are short snippets of text, 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. This proved to be a labor-intensive process. Furthermore, Xu et al. did not perform a large-scale assessment of the reliability of ROUGE for evaluating answers to "definition" questions (which we describe in Section 6). As explicated by C.-Y. Lin and Hovy (2003), among other researchers, metrics developed for one task cannot usually be directly applied to another task without modification. Our experiments show that ROUGE is less well-suited for evaluating answers to complex questions because it fails to capture the finer intricacies of the official scoring metric such as the vital/okay nugget distinction and precision length penalty.

In general, we have noted a recent convergence between multi-document summarization and question answering (Lin and Demner-Fushman, 2005). The move towards more complex information needs in question answering is complemented by the development of topic-focused summarization (Dang, 2005). Most notably, the DUC 2005 task requires systems to generate answers to natural language questions based on a collection of known relevant documents: "The system task in 2005 was to synthesize from a set of 25–50 documents a brief, well-organized, fluent answer to a need for information that cannot be met by just stating a name, date, quantity, etc." (DUC 2005 guidelines[1]). These guidelines were modeled after the *information synthesis* task suggested by Amigó et al. (2004), which they characterize as "the process of (given a complex information need) extracting, organizing, and inter-relating the pieces of information contained in a set of relevant documents, in order to obtain a comprehensive, non-redundant report that satisfies the information need". This trend represents a remarkable opportunity for cross-fertilization and intellectual dialog between two mostly-disjoint communities.

4 Algorithm Description

As previously mentioned, it has been widely assumed that matching nuggets from the assessors' answer key with systems' responses must be performed manually because it involves semantics (Voorhees, 2003). In actuality, however, this assumption has not been experimentally verified, and we 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. Matching longer n-grams (bigrams, trigrams, etc.) has proven useful in machine translation because it measures the fluency of system-generated responses, an important aspect of translation quality. However, since almost all current question answering systems employ extractive techniques (i.e., no natural language generation), fluency is not usually a concern. The presence or absence of relevant content, captured via unigram co-occurrence, is the more important factor in both document summarization and complex question answering.

[1] `http://www-nlpir.nist.gov/projects/duc/duc2005/tasks.html`

Testset	*okay*		*vital*		*total*	
	mean	σ	mean	σ	mean	σ
TREC 2003 "definition"	4.2	2.7	4.1	3.4	8.3	3.7
TREC 2004 "other"	5.4	2.1	3.6	1.8	9.1	2.8
TREC 2005 "other"	6.0	2.3	4.1	2.4	10.1	2.5

Table 2: Descriptive statistics (mean and σ) of the number of nuggets in each year's testset.

The idea behind POURPRE is relatively straightforward: calculate a match score for each nugget by summing the unigram co-occurrences between terms from the nugget and terms from the system response. We decided to start with the simplest possible approach: count the term 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 extracts of different 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 (see Figure 2), except that the automatically-derived 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. This remained exactly the same under POURPRE.

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 contains 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. The effects of stemming, i.e., matching stemmed terms derived from the Porter stemmer, were also examined.

Finally, we attempted two different methods for aggregating results: microaveraging and macroaveraging. For microaveraging, final per-run scores were calculated by computing 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. This is a distinction that is often made in evaluating text classification.

5 Validation Methodology

Submitted runs to the TREC 2003, 2004, and 2005 question answering tracks were used to validate POURPRE (2003 "definition" questions, 2004 and 2005 "other" questions). There were 50 questions in 2003 testset, 65 in 2004, and 75 in 2005. Table 2 shows the median and standard deviation of okay and vital nuggets in each year's testset. In total, there were 54 runs from TREC 2003, 63 from TREC 2004, and 72 from TREC 2005. For each experiment, system runs were automatically scored with a variant of our POURPRE metric, and the results were then correlated with the official scores generated by humans. As a baseline, we compared POURPRE to ROUGE, BLEU, and NIST, using a simple concatenation of all answer nuggets as the reference output.

Before presenting results, it is important to first define the basis for assessing the validity of an automatic evaluation metric. Direct correlation between official scores and automatically-generated scores, as measured by Pearson's r, seems like an obvious metric for quantifying the validity 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. 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; earlier uses of this metric for validation question answering evaluations can be seen in, for example, (Voorhees and Tice, 2000). 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. In this study, we primarily focus on Kendall's τ, but also report Pearson's r where appropriate.

6 Validation Results

The Kendall's τ correlations between rankings produced by POURPRE and the official rankings are shown in Table 3 for all testsets. The Pearson's r correlations between the two sets of scores are shown in Table 4. These two tables contain results of four separate POURPRE variants along two different parameters: scoring by term counts only vs. scoring by term *idf*, and microaveraging vs. macroaveraging. We present results for six distinct sets of runs: In TREC 2003, the value of β, which controlled the relative importance of precision and recall, was arbitrarily set to five, which was later determined to favor recall too heavily. As a result, β was readjusted to three in TREC 2004. In our experiments with TREC 2003, we report figures for both values. In the answer key to the TREC 2005 "other" questions, we noticed that many nuggets were written with shorthand or contained typos, e.g.,

> Bob Dole rcvd $45000 weekly while endorsement ran.
> CBS bdcsts event to 100+ countries.
> Purchasing 28 wud cost NZ $346 million (US).

From the assessors' point of view, answer nuggets serve as notes to assist in scoring; they were not aware of the nuggets' subsequent use as an answer key. In addition to using the verbatim answer key for the 2005 "other" questions, we also manually corrected it (without altering nugget content) and reran experiments. Finally, we also report results of experiments using the 2005 data that excludes a manually-generated run; this is denoted "auto only" in Tables 3 and 4 (more on this below).

Run	POURPRE			
	micro, cnt	macro, cnt	micro, *idf*	macro, *idf*
2003 "definition" ($\beta = 5$)	**.890**	.878	.860	.875
2003 "definition" ($\beta = 3$)	.846	**.886**	.848	.876
2004 "other" ($\beta = 3$)	.785	**.833**	.806	.812
2005 "other" ($\beta = 3$)	.598	**.709**	.679	.698
2005 "other" ($\beta = 3$, auto only)	.590	**.731**	.707	.725
2005 "other" ($\beta = 3$, corrected)	.572	**.710**	.683	.700

Table 3: Kendall's τ correlation between rankings generated by four different POURPRE variants and official scores on various TREC testsets.

Run	POURPRE			
	micro, cnt	macro, cnt	micro, *idf*	macro, *idf*
2003 "definition" ($\beta = 5$)	.977	**.982**	.978	**.982**
2003 "definition" ($\beta = 3$)	.958	**.981**	.970	.979
2004 "other" ($\beta = 3$)	.915	**.964**	.950	.956
2005 "other" ($\beta = 3$)	.849	**.916**	.894	.903
2005 "other" ($\beta = 3$, auto only)	.852	**.935**	.922	.929
2005 "other" ($\beta = 3$, corrected)	.843	**.916**	.894	.903

Table 4: Pearson's r correlation between scores generated by four different POURPRE variants and official scores on various TREC testsets.

Interestingly, scoring based on macroaveraged term counts outperformed any of the other variants—it does not appear that *idf* term weighting helps at all. Also surprising is the fact that correcting errors in the answer key had almost no effect on performance. This suggests that POURPRE is relatively insensitive to errors in the reference nuggets. Although the best variant of POURPRE, macroaveraging with term counts, predicts human preferences well for TREC 2003 and TREC 2004, the correlations are much lower for the 2005 testset (although the values are still high enough for POURPRE to be useful in guiding the development of future systems). In Section 7.4, we will examine this issue in more detail.

How does POURPRE stack up against using ROUGE[2], BLEU, and NIST[3] to directly evaluate answers to complex questions? For all three measures, the reference output consisted simply of all answer nuggets concatenated together. Results are shown in Table 5 for Kendall's τ and Table 6 for Pearson's r. We tested three variants of ROUGE: the default with neither stopword removal nor stemming (base), with stopword removal but no stemming (stop), and with both stopword removal and stemming (s+s). For BLEU and NIST, we report both the unigram score (uni) and total score (tot). In all cases except for the 2005 "other" questions with corrected nuggets, the POURPRE condition with macroaveraging and term counts outperforms all variants of the other three metrics (based on Kendall's τ). Scatter graphs plotting official F-scores against POURPRE scores (macro, count) and ROUGE scores (stopword removal, no stemming) for TREC 2003 ($\beta = 5$), TREC 2004 ($\beta = 3$), and TREC 2005 ($\beta = 3$) are shown in Figure 3. Corresponding graphs for other variants appear similar, and are not shown here.

In the scatter plot for TREC 2005 ($\beta = 3$), we note a very interesting data point on the rightmost edge (marked by two arrows): the highest-scoring run according to human assessors performs poorly with POURPRE. There is a straightforward explanation: answers contained in that run were

[2]We used ROUGE-1 recall values from ROUGE-1.4.2; experiments were also conducted on ROUGE-1 precision, ROUGE-2 precision, and ROUGE-2 recall. Empirically, ROUGE-1 performed the best.

[3]For both, we used the default parameter settings.

Run	POURPRE	ROUGE			BLEU		NIST	
	macro, cnt	base	stop	s+s	uni	tot	uni	tot
2003 "definition" ($\beta = 5$)	**.878**	.807	.843	.834	.154	.290	.192	.200
2003 "definition" ($\beta = 3$)	**.886**	.780	.816	.810	.169	.306	.207	.216
2004 "other" ($\beta = 3$)	**.833**	.780	.786	.771	.168	.409	.235	.235
2005 "other" ($\beta = 3$)	**.709**	.662	.670	.675	.213	.368	.226	.238
2005 "other" ($\beta = 3$, auto only)	**.731**	.700	.698	.704	.190	.349	.203	.215
2005 "other" ($\beta = 3$, corrected)	.709	.717	**.764**	.760	.213	.374	.228	.234

Table 5: Comparison of POURPRE and other metrics in terms of Kendall's τ. For ROUGE: base=no stemming, no stopword removal; stop=no stemming, stopword removal,s+s=stemming and stopword removal. For BLEU/NIST: uni=unigram scoring, tot=total.

Run	POURPRE	ROUGE			BLEU		NIST	
	macro, cnt	base	stop	s+s	uni	tot	uni	tot
2003 "definition" ($\beta = 5$)	**.982**	.958	.964	.962	.159	.278	.215	.224
2003 "definition" ($\beta = 3$)	**.981**	.936	.942	.938	.224	.344	.276	.285
2004 "other" ($\beta = 3$)	**.964**	.924	.933	.931	.210	.539	.293	.304
2005 "other" ($\beta = 3$)	**.916**	.900	.902	.902	.416	.584	.479	.491
2005 "other" ($\beta = 3$, auto only)	.935	**.940**	.938	**.940**	.351	.605	.424	.438
2005 "other" ($\beta = 3$, corrected)	.916	.906	**.930**	.928	.416	.586	.478	.490

Table 6: Comparison of POURPRE and other metrics in terms of Pearson's r. For ROUGE: base=no stemming, no stopword removal; stop=no stemming, stopword removal,s+s=stemming and stopword removal. For BLEU/NIST: uni=unigram scoring, tot=total.

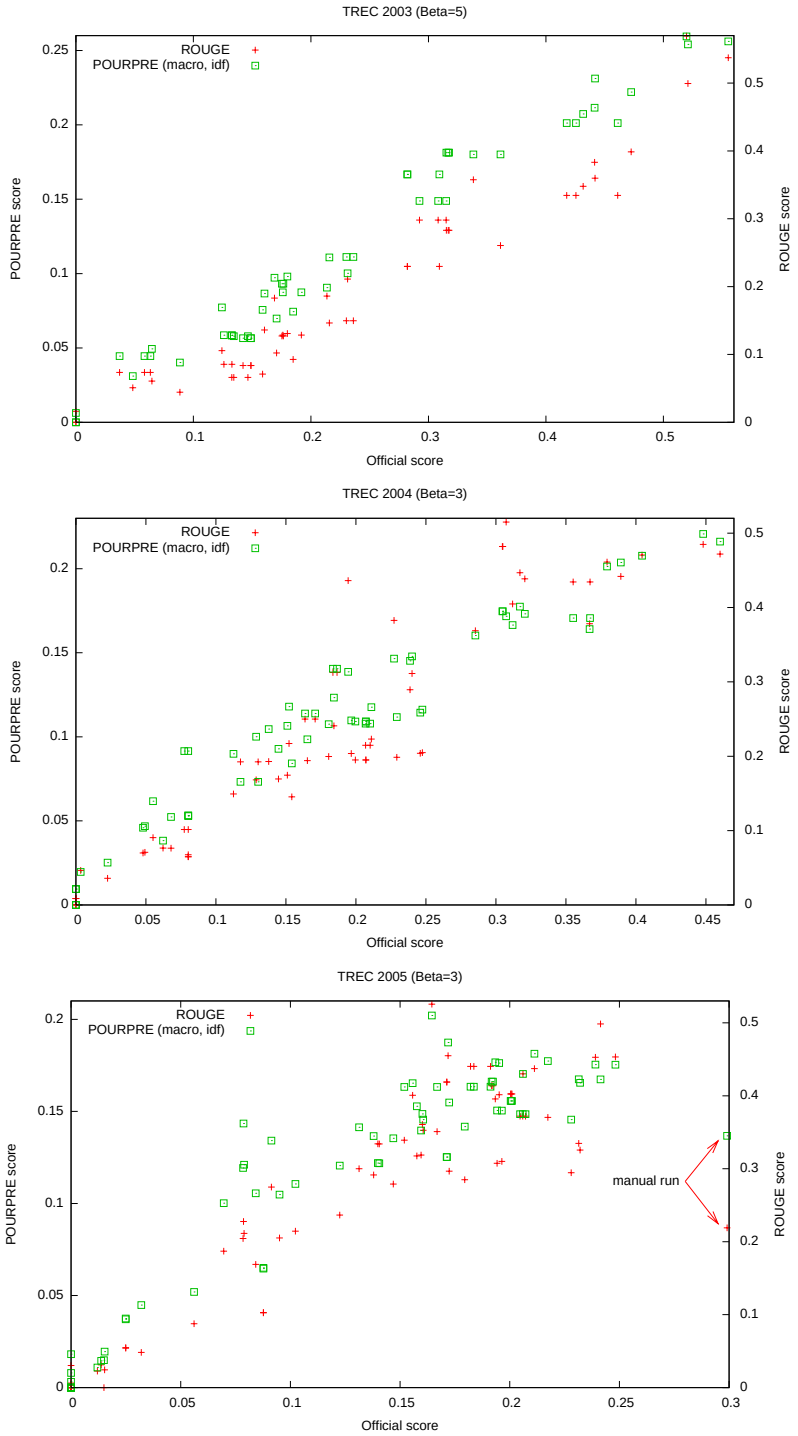

Figure 3: Scatter graph of official scores plotted against POURPRE (macro, count) and ROUGE scores (stopword removal, no stemming) for TREC 2003 ($\beta = 5$), TREC 2004 ($\beta = 3$), TREC 2005 ($\beta = 3$).

Run	unstemmed	stemmed
2003 "definition" ($\beta = 5$)	0.878	0.895
2003 "definition" ($\beta = 3$)	0.886	0.897
2004 "other" ($\beta = 3$)	0.833	0.825
2005 "other" ($\beta = 3$)	0.709	0.680
2005 "other" ($\beta = 3$, auto only)	0.731	0.703
2005 "other" ($\beta = 3$, corrected)	0.710	0.685

Table 7: The effect of stemming on Kendall's τ; all runs with the (macro, count) variant of POURPRE.

actually supplied by a trained librarian, in an attempt to better understand human performance on this particular task (Lin et al., 2005). The human run did not primarily consist of sentence extracts, but truly summarized information nuggets within the documents—and thus was not fairly assessed by POURPRE, which was designed (like ROUGE) for extractive responses. Experiments with this run removed are denoted as "auto only" in Tables 3 through 6. It can be seen that removing this manual run increases both Kendall's τ and Pearson's r correlations.

The effect of stemming on Kendall's τ and Pearson's r between POURPRE (macro, count) and official scores is shown in Table 7. Results from the same stemming experiment on the other POURPRE variants are similarly inconclusive.

Why does neither stemming nor *idf*-weighting improve the correlation of POURPRE with official TREC results? After detailed examination of the data, we were able to produce explanations for both seemingly counter-intuitive results.

Stemming is primarily a recall-focused technique that handles morphological divergences between the reference nuggets and system outputs—it allows terms to match despite differences in surface morphology; cf. (Kraaij and Pohlmann, 1996). Reference nuggets can be broadly classified into two categories: those that are essentially extracts of documents from the corpus, and more "abstractive" ones that summarize content found in multiple documents. For extractive nuggets, stemming doesn't make a difference because most TREC systems are extractive also, taking passages directly from relevant documents. Hence, terms already match without additional morphological processing. For "abstractive" nuggets, since they are often paraphrases of system responses, stemming doesn't help bridge the gap between the semantically equivalent, but superficially different, forms anyway. Thus, stemming has a relatively minor overall effect on the performance of POURPRE.

A term-weighting scheme based on *idf* doesn't affect performance much for similar reasons. The rationale behind *idf* is that more emphasis would be placed on content words, so that matching frequently-occurring words would be penalized by comparison. However, the reference nuggets in general have a high proportion of content words (since they must be sufficiently descriptive to assist in the evaluation process), which lessens the impact of *idf*-weighting.

Building on the basic results reported here, we describe in the next section several additional experiments aimed at a better understanding of performance factors affecting POURPRE.

7 Understanding Performance Effects

Our experimental results show a strong correlation between official human-generated TREC scores and automatically-computed POURPRE scores, indicating that a simple term-matching approach can capture various aspects of the evaluation process. Starting from a set of questions and a list of relevant nuggets, POURPRE can accurately assess the performance of a complex question answering system without any human intervention.

To better understand the limitations of our work, it is worth mentioning that POURPRE, or any automatic evaluation metric based on substring overlap, for that matter, cannot actually replace human judgments or capture all the fine-grained semantic considerations that go into matching system responses with reference output. We have merely demonstrated that it is possible to approximate human judgments in a manner that facilitates system development. It is certainly possible that at some future point in time, POURPRE will "outlive its usefulness", for exactly the reasons stated above. Nevertheless, it is our hope that the metric will facilitate the creation of more sophisticated systems until then. Similar limitations exist for all automatic evaluation metrics based solely on surface string matching. The machine translation community is contending with exactly such an issue: it is probable in the not–so–distant future that system performance will be indistinguishable from human performance according to BLEU. At that point, the MT community would need to develop metrics that are better able to capture various linguistic phenomena (e.g., paraphrases). Meanwhile, however, BLEU will remain useful in guiding system development.

Although POURPRE addresses the problem of automatically assessing system output, it does not assist in the creation of test collections themselves, which is a time- and resource-consuming endeavour. In particular, our metric depends on the existence of reference nuggets, which, in the TREC setup, is "distilled" by a human after examining all system results (and performing additional topic research). Similar limitations apply to all metrics of this sort—BLEU requires human-generated reference translations and ROUGE requires human-generated summaries. Naturally, since the notion of relevance features prominently in the evaluation of question answering systems, and since relevance ultimately boils down to human judgments, there are theoretical limitations on the degree to which evaluations can be automated. However, it may be possible to develop systems that assist in the process of nugget creation; for example, the work on Basic Elements (Hovy et al., 2005) represents such an attempt.

The results reported in the previous section leads to a number of interesting questions:

- Why exactly does POURPRE work better than ROUGE?
- How reliable are POURPRE scores and what is the margin of error associated with them?
- How sensitive is POURPRE to different precision–recall tradeoff points?
- How does the vital/okay distinction affect the validity of POURPRE scores?

In the following subsections, we describe a series of detailed experiments that address the above issues. The goal is to gain more insight into the nugget evaluation methodology and the automatic evaluation process.

7.1 Algorithm Analysis and Hybridized Matching

We believe that POURPRE better correlates with official scores than ROUGE 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, i.e., it can reveal *why* an answer received a particular 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.

Comparisons with the baselines of ROUGE and BLEU/NIST support our hypothesis concerning the nature of complex question answering. The correlation between ROUGE and official human-generated scores is much higher than that of BLEU or NIST—conceptually, the task of answering complex questions is closer to document summarization than it is to machine translation, in that both are recall-oriented

and focus primarily on content. Since almost all question answering systems employ extractive techniques, fluency (i.e., precision) is not usually an issue. However, as systems begin to employ abstractive techniques to generate answers, evaluation metrics will need start assessing answer fluency; however, such systems appear to be just beyond the short-term horizon.

To better understand why POURPRE outperforms ROUGE, we conceptually decomposed the evaluation of complex questions into two distinct stages. The first deals with the matching of the nuggets themselves, while the second deals with the manner in which the nugget match scores are combined to generate a final F-score for a particular question. Since POURPRE attempts to capture both stages of the evaluation process, it outperforms ROUGE, which does not explicitly match individual nuggets, nor does it capture the TREC scoring model.

In order to confirm this hypothesis, we experimented with a POURPRE–ROUGE hybrid that matches individual nuggets using ROUGE, but combines the match scores in a manner consistent with the TREC scoring model, like POURPRE. Nugget recall and length allowance, the two variables that determine F-score, can be computed in the following manner:[4]

- **Nugget recall.** Instead of lumping all nuggets in a single "summary", we treated each individual vital nugget as a distinct reference summary and scored system output using ROUGE in the following manner:

$$\mathcal{R} = \frac{\sum_{s \in A} \text{ROUGE}_{\text{vital}}(s)}{R} \tag{1}$$

  Where A is the set of answer passages returned by a system. Since ROUGE output is bounded between zero and one, this produced the equivalent of nugget recall, i.e., r/R in Figure 2.

- **Length allowance.** Length allowance is a product of a constant (100 non-whitespace characters) and the number of vital *and* okay nuggets matched. With ROUGE, it can be computed in the following manner (using all vital and okay nuggets as separate reference summaries):

$$\alpha = 100 \times \sum_{s \in A} \text{ROUGE}_{\text{vital+okay}}(s) \tag{2}$$

  Once again, A represents the set of all answer passages returned by a system. This produces the equivalent of $r + a$ in Figure 2.

How does this POURPRE–ROUGE hybrid (which we term POURPRE-R) stack up against POURPRE alone? Results of correlation experiments are shown in Table 8 (for Kendall's τ) and Table 9 (for Pearson's r). For the ROUGE component inside POURPRE-R, we experiment with three variants of ROUGE-1 recall scores: the default with neither stopword removal nor stemming (base), with stopword removal but no stemming (stop), and with both stopword removal and stemming (s+s). Correlation values are juxtaposed with the best POURPRE and ROUGE values from Tables 5 and 6.

Our experiments show that POURPRE-R is comparable to the simpler version of POURPRE, and in some cases, slightly better. This confirms our hypothesis regarding why POURPRE achieves higher correlations. Of the two stages in the evaluation process, it does not appear that the nugget matching process plays a particularly important role—the choice of the maximal unigram recall method employed by POURPRE or the use of multiple summaries with ROUGE inside POURPRE-R does not appear to have a large impact on performance. The contribution to correlation appears to be coming from explicit recognition of the nugget-based evaluation methodology.

[4]We are indebted to an anonymous reviewer for suggesting this experiment.

Run	POURPRE	ROUGE	POURPRE-R		
	macro, cnt	stop	base	stop	s+s
2003 "definition" ($\beta = 5$)	.878	.843	.866	.877	.890
2003 "definition" ($\beta = 3$)	.886	.816	.864	.880	.897
2004 "other" ($\beta = 3$)	.833	.786	.808	.837	.855
2005 "other" ($\beta = 3$)	.709	.670	.708	.692	.751
2005 "other" ($\beta = 3$, auto only)	.731	.698	.728	.711	.769
2005 "other" ($\beta = 3$, corrected)	.709	.764	.713	.767	.756

Table 8: Comparison of Kendall's τ correlation for POURPRE-R (base=no stemming, no stopword removal; stop=no stemming, stopword removal,s+s=stemming and stopword removal).

Run	POURPRE	ROUGE	POURPRE-R		
	macro, cnt	stop	base	stop	s+s
2003 "definition" ($\beta = 5$)	.982	.964	.977	.981	.983
2003 "definition" ($\beta = 3$)	.981	.942	.975	.980	.982
2004 "other" ($\beta = 3$)	.964	.933	.948	.972	.970
2005 "other" ($\beta = 3$)	.916	.902	.905	.907	.920
2005 "other" ($\beta = 3$, auto only)	.935	.938	.924	.926	.938
2005 "other" ($\beta = 3$, corrected)	.916	.930	.905	.926	.920

Table 9: Comparison of Pearson's r correlation for POURPRE-R (base=no stemming, no stopword removal; stop=no stemming, stopword removal,s+s=stemming and stopword removal).

7.2 Rank Swap Analysis

What do rank correlations mean in the context of real-world evaluations? An analysis of rank swaps can help us better interpret the effectiveness of an evaluation metric, and is a common technique used in information retrieval research. A rank swap is said to have occurred if the relative ranking of two runs is different under different scoring conditions, i.e., according to one condition, system A is better than system B, but the opposite conclusion is drawn based on the other condition. Rank swaps are significant because they may prevent researchers from confidently drawing conclusions about the relative effectiveness of different techniques or systems. For example, observed rank swaps between official and POURPRE scores indicate cases where the automatic scoring metric is not accurately capturing human judgments.

We analyzed rank swaps between official human-generated scores and POURPRE scores on testsets from TREC 2003 ($\beta = 5$), TREC 2004 ($\beta = 3$), and TREC 2005 ($\beta = 3$). For the 2003 testset, we observed 81 rank swaps out of a total of 1431 pairwise comparisons for 54 runs (5.7% of all comparisons). For the 2004 testset, we observed 157 rank swaps out of a total of 1953 pairwise comparisons for 63 runs (8.0% of all comparisons). For the 2005 testset, we observed 357 rank swaps out of a total of 2556 pairwise comparisons for 72 runs (14% of all comparisons). Removing the human-generated run reduces the number of rank swaps to 320.

Histograms of these rank swaps, binned by the difference in official score between the two runs, are shown in Figure 4 for the four different testsets described above. As can be seen, on the TREC 2003 ($\beta = 5$) testset, 48 rank swaps (59.3%) 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 0.061. The other histograms can be interpreted similarly.

What is the meaning of these rank swap analyses? Since measurement error is an inescapable fact

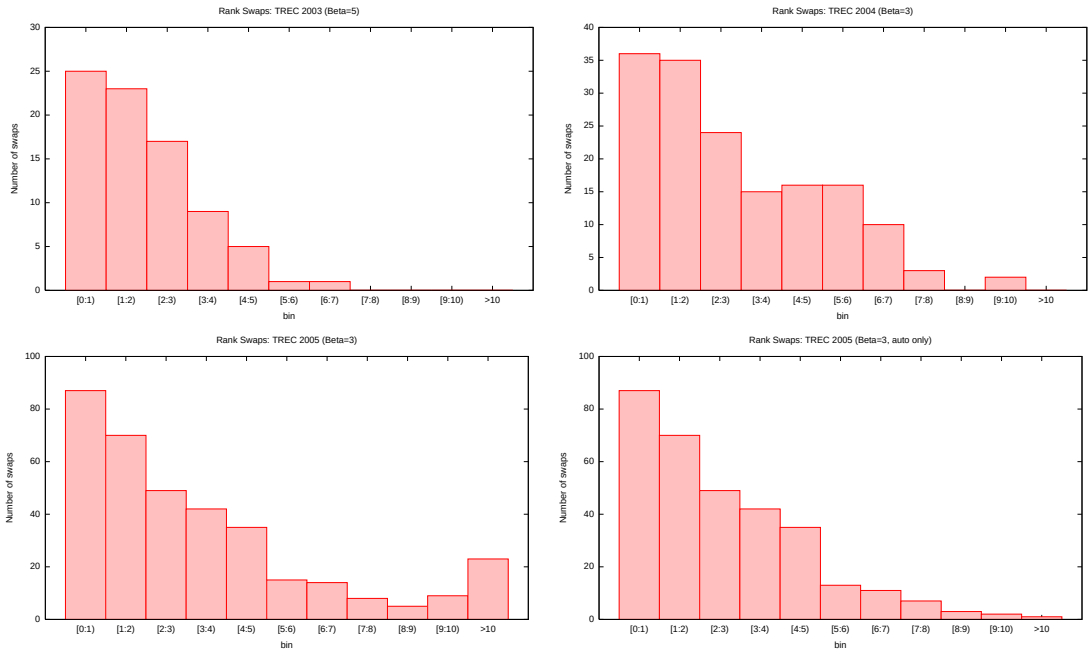

Figure 4: Histogram of rank swaps binned by differences in official score: TREC 2003 ($\beta = 5$) (top left), TREC 2004 ($\beta = 3$) (top right), and TREC 2005 ($\beta = 3$) (bottom left), TREC 2005 ($\beta = 3$, auto only) (bottom right). Bin units are in hundredths.

of evaluation, we need not be concerned with rank swaps for which differences in the original score is less than the bounds of error for the evaluation (since one could not distinguish the two runs with confidence anyway). 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. The histogram for TREC 2004 shows two rank swaps in range of [0.09, 0.10), but none above .1; the 2005 testset shows greater variability, with 23 rank swaps above that threshold (although many of which can be attributed to the top-scoring run, which was generated by a human).

From these results, we can conclude that evaluation of "definition" questions is relatively coarse-grained. In general, POURPRE accurately predicts human judgments on the quality of answers to complex questions, to the extent where the margin of error is approximately equal to the measurement uncertainty associated with the evaluation itself.

7.3 Balancing Precision and Recall

The final F-score of an answer is the harmonic mean of nugget recall and nugget precision, where relative weight between the two components is controlled by the β parameter. Obviously, a single measure of effectiveness is desirable from an evaluation point of view, but it is important to note that the specific setting of β operationalizes one particular tradeoff between recall and precision. For different applications, the relative importance of these two components may vary dramatically. The model of complex question answering implemented at TREC places significantly more emphasis on recall, with β set to five in 2003 and three the years thereafter.

Recognizing that different parameter settings may be appropriate for other types of applications, we ran our experiments with β as the independent variable. More specifically, we computed the Kendall's τ correlation between POURPRE scores (macroaveraging and term counts) and official scores under

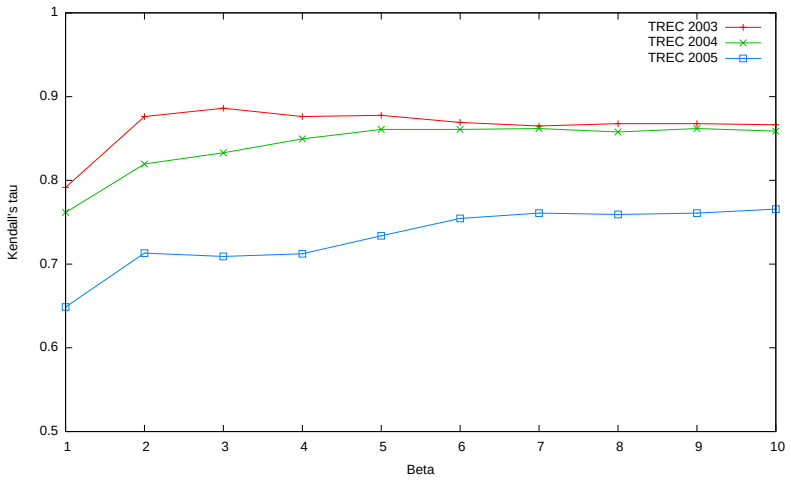

Figure 5: Kendall's τ for varying values of β.

Testset	# q's	1 vital	2 vital
TREC 2003	50	3	10
TREC 2004	64	2	15
TREC 2005	75	5	16

Table 10: Number of questions with few vital nuggets in the different testsets.

various settings of β. These results are shown in Figure 5.

We observe that an emphasis on recall yields better correlation between POURPRE and official scores. A β value of one (equal precision and recall) results in a lower Kendall's τ value than a β value of two, but higher values of β don't seem to make much of a difference. In general, these results confirm that POURPRE is primarily a recall-oriented metric, much like ROUGE. This makes sense because unigram overlap measures the amount of shared content between system and reference nuggets. Precision enters into the calculation only as a length penalty.

7.4 Implications of the Vital/Okay Distinction

The vital/okay distinction on nugget labels attempts to capture the intuition that some facts are more important than others. Nevertheless, this binary distinction may be problematic from an automatic evaluation point of view. In this section, we explore the implications of this evaluation aspect in greater detail.

From Tables 3 and 4, it can be seen that the accuracy with which POURPRE predicts human judgments varies from testset to testset. In particular, correlation is highest for the TREC 2003 testset, slightly lower for the TREC 2004 testset, and noticeably lower for the TREC 2005 testset.

We believe that this can be attributed to the distribution of vital and okay nuggets in the three testsets. Table 10 shows statistics about the number of questions that have only one or two vital nuggets. Compared to the size of the testset, these numbers are relatively large. As a concrete example, "F16" is the target for question 71.7 from TREC 2005. The only vital nugget is "First F16s built in 1974". With only one vital nugget, a system's score on that particular question would be zero unless it returned the sole vital nugget, since okay nuggets figured only in the precision calculation—in other words, a system had only one chance to achieve a non-zero score. This results in highly unpredictable scores that were often dictated by chance as much as by actual system performance. The same effect

Run	POURPRE	ROUGE
2003 "definition" ($\beta = 5$)	0.910	0.886
2003 "definition" ($\beta = 3$)	0.915	0.872
2004 "other" ($\beta = 3$)	0.868	0.840
2005 "other" ($\beta = 3$)	0.807	0.789

Table 11: Kendall's τ correlation for POURPRE and ROUGE (stopword removal, no stemming) based on variant answer key where all nuggets are considered vital.

is also present for questions with only two vital nuggets, although to a lesser degree. This phenomenon is further amplified by the final F-score's emphasis on recall over precision. As a result, POURPRE is not able predict human scores on the TREC 2005 testset as well.

To test this hypothesis, we created a variant answer key in which all nuggets were considered vital and reran our experiments. Because assessors recorded nugget matches independent of vital/okay status, it is possible to recalculate scores based on different conditions. These results (Kendall's τ) are shown in Table 11; for brevity, only the best POURPRE (macro, count) and ROUGE (stopword removal, no stemming) configurations are shown. Indeed, higher correlation is observed when all nuggets are considered vital, with POURPRE still outperforming ROUGE. Nevertheless, there is still a gap in performance between the three different testset, which can be attributed to the quality of the different reference nuggets, as discussed in Section 6.

The vital/okay distinction aims to capture differences in assessors' notions of relevance. As with many other information retrieval tasks, legitimate differences in opinion about relevance are an inescapable fact of evaluating definition/other questions—systems are designed to satisfy real-world information needs, and users inevitably disagree on which 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 (across a broad sampling of users)? 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 the remainder of this section, we explore the effect of judgment variability on the stability and reliability of TREC nugget-based evaluation methodology.

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 question, for 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. Consider some relevant nuggets for the question "What is Bausch & Lomb?":

> world's largest eye care company
> about 12000 employees
> in 50 countries
> approx. $1.8 billion annual revenue
> based in Rochester, New York

According to the official assessment, the first four nuggets are vital and the fifth is not. This means that the location of Bausch & Lomb's headquarters is considered less important than employee count and revenue. Such mysteries are common within answer nuggets across all examined testsets, and the vital/okay distinction does not seem to follow from any deducible principle. As a result, there is little hope for systems to learn and exploit this difference. Without any guiding principles, how can we expect our systems to focus more on returning vital nuggets?

Run	everything vital	vital/okay flipped	random judgments
TREC 2003 ($\beta = 5$)	0.927	0.802	0.822 ± 0.042
TREC 2003 ($\beta = 3$)	0.920	0.796	0.808 ± 0.043
TREC 2004 ($\beta = 3$)	0.919	0.859	0.841 ± 0.038
TREC 2005 ($\beta = 3$)	0.854	0.766	0.777 ± 0.043

Table 12: Kendall's τ correlations between the official scores and scores under different variations of judgments and the official scores. The 95% confidence interval is presented for the random judgments case.

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 12. Note that this experiment was conducted with the manually-evaluated system responses, not our POURPRE metric. 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 are reported.

These results suggest that system rankings are sensitive to assessors' 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).

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 keys (with the exception of "everything vital"). This means that further refinement of the POURPRE metric to increase correlation with human-generated scores may not be particularly meaningful; it might essentially amount to 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. Recently, we have looked into techniques for combining judgments from multiple assessors in order to obtain a more refined estimate of nugget importance (Lin and Demner-Fushman, 2006). This was inspired by the pyramid scheme (Nenkova and Passonneau, 2004), related work from the multi-document summarization literature.

8 Conclusion

The nugget-based evaluation methodology described in this paper is broadly applicable to many types of complex questions—not only the definition/other questions specifically explored here, but also relationship and opinion questions that are beginning to benefit from large-scale formal evaluations at TREC. Our work shows that POURPRE can be employed to automatically evaluate answers to complex questions, and that rankings generated by this method correlate well with human-generated rankings. We hope that POURPRE can accomplish for complex 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.[5]

9 Acknowledgments

This work was supported in part by ARDA's Advanced Question Answering for Intelligence (AQUAINT) Program. This article is an expanded version of a paper originally published in the Proceedings of the 2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005). The first author would like to thank Esther and Kiri for their loving support.

References

Enrique Amigó, Julio Gonzalo, Victor Peinado, Anselmo Peñas, and Felisa Verdejo. 2004. An empirical study of information synthesis task. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*.

Bogdan Babych and Anthony Hartley. 2004. Extending the BLEU MT evaluation method with frequency weightings. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.

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

Hoa Dang. 2005. Overview of DUC 2005. In *Proceedings of the 2005 Document Understanding Conference (DUC 2005) at NLT/EMNLP 2005*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In *Proceeding of 2002 Human Language Technology Conference (HLT 2002)*.

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)*.

Eduard Hovy, Chin-Yew Lin, and Liang Zhou. 2005. Evaluating DUC 2005 using Basic Elements. In *Proceedings of the 2005 Document Understanding Conference (DUC 2005) at NLT/EMNLP 2005*.

Wessel Kraaij and Renée Pohlmann. 1996. Viewing stemming as recall enhancement. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1996)*.

Alex Kulesza and Stuart M. Shieber. 2004. A learning approach to improving sentence level MT evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*.

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

Jimmy Lin and Dina Demner-Fushman. 2005. 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.*

Jimmy Lin and Dina Demner-Fushman. 2006. Will pyramids built of nuggets topple over? In *Proceedings of the 2006 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2006).*

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).*

Chin-Yew Lin and Franz Josef Och. 2004. ORANGE: A method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004).*

Jimmy Lin, Eileen Abels, Dina Demner-Fushman, Douglas W. Oard, Philip Wu, and Yejun Wu. 2005. A menagerie of tracks at Maryland: HARD, Enterprise, QA, and Genomics, oh my! In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005).*

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).*

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002).*

Radu Soricut and Eric Brill. 2004. A unified framework for automatic evaluation using n-gram co-occurrence statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004).*

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000).*

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 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).*

Jinxi Xu, Ralph Weischedel, and Ana Licuanan. 2004. Evaluation of an extraction-based approach to answering definition questions. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004).*