Evaluating Answers to Definition Questions

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, MD 20899
ellen.voorhees@nist.gov

Abstract

This paper describes an initial evaluation of systems that answer questions seeking definitions. The results suggest that humans agree sufficiently as to what the basic concepts that should be included in the definition of a particular subject are to permit the computation of concept recall. Computing concept precision is more problematic, however. Using the length in characters of a definition is a crude approximation to concept precision that is nonetheless sufficient to correlate with humans' subjective assessment of definition quality.

The TREC question answering track has sponsored a series of evaluations of systems' abilities to answer closed class questions in many domains (Voorhees, 2001). Closed class questions are fact-based, short answer questions. The evaluation of QA systems for closed class questions is relatively simple because a response to such a question can be meaningfully judged on a binary scale of right/wrong. Increasing the complexity of the question type even slightly significantly increases the difficulty of the evaluation because partial credit for responses must then be accommodated.

The ARDA AQUAINT[1] program is a research initiative sponsored by the U.S. Department of Defense aimed at increasing the kinds and difficulty of the questions automatic systems can answer. A series of pilot evaluations has been planned as part of the research agenda of the AQUAINT program. The purpose of each pilot is to develop an effective evaluation methodology for systems that answer a certain kind of question. One of the first pilots to be implemented was the Definitions Pilot, a pilot to develop an evaluation methodology for questions such as *What is mold?* and *Who is Colin Powell?*.

[1]See `http:///www.ic-arda.org/InfoExploit/aquaint/index.html`.

This paper presents the results of the pilot evaluation. The pilot demonstrated that human assessors generally agree on the concepts that should appear in the definition for a particular subject, and can find those concepts in the systems' responses. Such judgments support the computation of concept recall, but do not support concept precision since it is not feasible to enumerate *all* concepts contained within a system response. Instead, the length of a response is used to approximate concept precision. An F-measure score combining concept recall and length is used as the final metric for a response. Systems ranked by average F score correlate well with assessors' subjective opinions as to definition quality.

1 The Task

The systems' task in the pilot was as follows. For each of 25 questions the system retrieved a list of text fragments such that each fragment was a component of the definition. The list was assumed to be ordered such that the more important elements in the definition appeared earlier in the list. There were no limits placed on either the length of an individual fragment or on the number of items in a list, though systems knew they would be penalized for retrieving extraneous information. Six AQUAINT contractors submitted eight runs to the pilot. The eight runs are labeled A–H in the discussion below.

The questions were developed by NIST assessors who searched a set of news articles for definition targets. The result of question development was a question phrased as either "Who is…" or "What is…", plus their own definition of the target. In general, these definitions consisted of one or two paragraphs of English prose.

2 Assessing System Responses

Each system response was independently judged by two different assessors. In what follows, the "author" assessor is the assessor who originally created the question; the

"other" assessor is the second assessor to judge the question. Each assessor performed two rounds of assessing per question.

In the first round of assessing, the assessor assigned two scores to the response from a system. One score was for the content of the response and the other for its organization, with each score on a scale of 0–10. A high content score indicated that the response contained most of the information it should contain and little misleading information. A high organization score indicated the response ranked the more important information before the less important information, and contained little or no irrelevant information. The final score for a question was a function of the organization and content scores, with the content score receiving much more emphasis.

The ranking of systems when using the question author to assign scores was FADEBGCH; the ranking was FAEGDBHC when using scores assigned by the other assessor. The final scores for the systems varied across assessors largely due to different interpretations of the organization score. Different assessors used different default scores when there was only one entry in the system response; organization scores also appeared to be strongly correlated with content scores. Despite these differences, the judgments do provide some guidance as to how a more quantitative scoring metric should rank systems. The assessors preferred the responses from system F over those from system A, which in turn was preferred over the remainder of the systems. Responses from systems C and H were the least preferred.

The goal of the second round of assessing was to support a more quantitative evaluation of the system responses. In this round of assessing, an assessor first created a list of "information nuggets" about the target using all the system responses and the question author's definition. An information nugget was defined as a fact for which the assessor could make a binary decision as to whether a response contained the nugget. The assessor then decided which nuggets were vital—nuggets that must appear in a definition for that definition to be good. Finally, the assessor went through each of the system responses and marked where each nugget appeared in the response. If a system returned a particular nugget more than once, it was marked only once.

Figure 1 shows an example of how one response was judged for the question *Who is Christopher Reeve?*. The left side of the figure shows the concept list developed by the assessor, with vital concepts marked with a star. The right side of the figure shows a system response with the concepts underlined and tagged with the concept number.

In Figure 1, each list entry has at most one concept marked. However, that was not generally the case. Many list entries contained multiple concepts while others contained none. Thus, using the list entry as the unit for evaluation is not sensible. Instead, we should calculate measures in terms of the concepts themselves. Computing concept recall is straightforward given these judgments; it is the ratio of the number of correct concepts retrieved to the number of concepts in the assessor's list. But the corresponding measure of concept precision, the ratio of the number of correct concepts retrieved to the total number of concepts retrieved, is problematic since the correct value for the denominator is unknown. A trial evaluation prior to the pilot showed that assessors found enumerating *all* concepts represented in a response to be so difficult as to be unworkable. For example, how many concepts are contained in "stars on Sunday in ABC's remake of 'Rear Window"? Using only concept recall as the final score is not workable either, since systems would not be rewarded for being selective: retrieving the entire document collection would get a perfect score for every question.

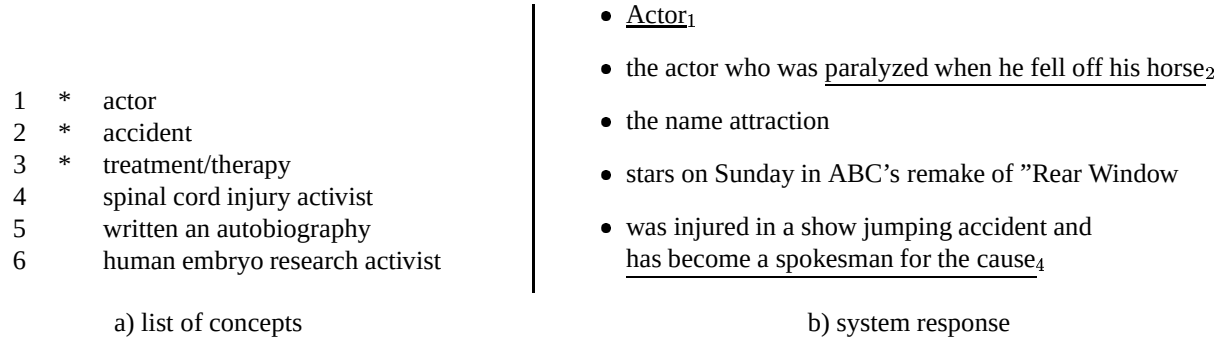
Borrowing from the evaluation of summarization systems (Harman and Over, 2002), we can use length as a (crude) approximation to precision. A length-based measure captures the intuition that users would prefer the shorter of two definitions that contain the same concepts. The length-based measure used in the pilot gives a system an allowance of 100 (non-white-space) characters for each correct concept it retrieves. The precision score is set to one if the response is no longer than this allowance. If the response is longer than the allowance, the precision score is downgraded using the function $\text{precision} = 1 - \frac{\text{length}-\text{allowance}}{\text{length}}$.

Remember that the assessors marked some concepts as vital and the remainder are not vital. The non-vital concepts act as a "don't care" condition. That is, systems should be penalized for not retrieving vital concepts, and penalized for retrieving items that are not on the assessor's concept list at all, but should be neither penalized nor rewarded for retrieving a non-vital concept. To implement the don't care condition, concept recall is computed only over vital concepts, while the character allowance in the precision computation is based on both vital and non-vital concepts. The recall for the example in Figure 1 is thus 2/3, and the character allowance is 300.

The final score for a response was computed using the F-measure, a function of both recall (R) and precision (P). The general version of the F-measure is

$$\mathrm{F} = \frac{(\beta^2 + 1)\mathrm{RP}}{\beta^2\mathrm{P} + \mathrm{R}}$$

where β is a parameter signifying the relative importance of recall and precision. The main evaluation in the pilot used a value of 5, indicating that recall is 5 times as important as precision. The value of 5 is arbitrary, but reflects both the emphasis given to content in the first round



a) list of concepts b) system response

Figure 1: Assessor annotation of a sample response for *Who is Christopher Reeve?*

author		other	
F	0.688	F	0.757
A	0.606	A	0.687
D	0.568	G	0.671
G	0.562	D	0.669
E	0.555	E	0.657
B	0.467	B	0.522
C	0.349	C	0.384
H	0.330	H	0.365

Table 1: Average F scores per system per assessor type.

of assessing and acknowledges the crudeness of the precision approximation.

Table 1 gives the average F scores for the pilot runs as evaluated using both assessors' judgments. As can be seen from the table, the rankings of systems are stable across different assessors in that the only difference in the rankings are for two runs whose scores are extremely similar (D and G). While the absolute value of the scores is different when using different assessors, the magnitude of the difference between scores is generally preserved. For example, there is a large gap between the scores for systems F and A, and a much smaller gap for systems C and H. The rankings also obey the ordering constraints suggested by the first round of assessing.

The different systems in the pilot took different approaches to producing their definitions. System H always returned a single text snippet as a definition. System B returned a set of complete sentences. System G tended to be relatively terse, while F and A were more verbose. The average length of a response for each system is A: 1121.2, B: 1236.5, C: 84.7, D: 281.8, E: 533.9, F: 935.6, G: 164.5, and H: 33.7. The differences in the systems are reflected in their relative scores when different β values are used. For example, when evaluated using $\beta = 2$ and the authors' judgments, the system ranking is GFDAECHB; for $\beta = 1$ the ranking is GDFAECHB. Thus as expected, as precision gains in importance, system G rises in the rankings, system B falls quickly, and system F also sinks.

3 Conclusion

The AQUAINT pilot evaluations are designed to explore the issues surrounding new evaluation methodologies for question answering systems using a small set of systems. If a pilot is successful, the evaluation will be transferred to the much larger TREC QA track. The definition pilot demonstrated that relative F scores based on concept recall and adjusted response length are stable when computed using different human assessor judgments, and reflect intuitive judgments of quality. The main measure used in the pilot strongly emphasized recall, but varying the F measure's β parameter allows different user preferences to be accommodated as expected. Definition questions will be included as a part of the TREC 2003 QA track where they will be evaluated using this methodology.

Acknowledgments

This paper was informed by the discussion that took place on the AQUAINT definition question mailing list. My thanks to the AQUAINT contractors who submitted results to the pilot evaluation, and especially to Ralph Weischedel and Dan Moldovan who coordinated the definition pilot.

References

Donna Harman and Paul Over. 2002. The DUC summarization evaluations. In *Proceedings of the International Conference on Human Language Technology*.

Ellen M. Voorhees. 2001. The TREC question answering track. *Journal of Natural Language Engineering*, 7(4):361–378.