Improving Search Results Quality by Customizing Summary Lengths

Michael Kaisser
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW
m.kaisser@sms.ed.ac.uk

Marti A. Hearst
UC Berkeley
102 South Hall
Berkeley, CA 94705
hearst@ischool.berkeley.edu

John B. Lowe
Powerset, Inc.
475 Brannan St.
San Francisco, CA 94107
johnblowe@gmail.com

Abstract

Web search engines today typically show results as a list of titles and short snippets that summarize how the retrieved documents are related to the query. However, recent research suggests that longer summaries can be preferable for certain types of queries. This paper presents empirical evidence that judges can predict appropriate search result summary lengths, and that perceptions of search result quality can be affected by varying these result lengths. These findings have important implications for search results presentation, especially for natural language queries.

1 Introduction

Search results listings on the web have become standardized as a list of information summarizing the retrieved documents. This summary information is often referred to as the document's *surrogate* (Marchionini et al., 2008).

In older search systems, such as those used in news and legal search, the document surrogate typically consisted of the title and important metadata, such as date, author, source, and length of the article, as well as the document's manually written abstract. In most cases, the full text content of the document was not available to the search engine and so no extracts could be made.

In web search, document surrogates typically show the web page's title, a URL, and information extracted from the full text contents of the document. This latter part is referred to by several different names, including *summary*, *abstract*, *extract*, and *snippet*. Today it is standard for web search engines to show these summaries as one or two lines of text, often with ellipses separating sentence fragments. However, there is evidence that the ideal result length is often longer than the standard snippet length, and that furthermore, result length depends on the type of answer being sought.

In this paper, we systematically examine the question of search result length preference, comparing different result lengths for different query types. We find evidence that desired answer length is sensitive to query type, and that for some queries longer answers are judged to be of higher quality.

In the following sections we summarize the related work on result length variation and on query topic classification. We then describe two studies. In the first, judges examined queries and made predictions about the expected answer types and the ideal answer lengths. In the second study, judges rated answers of different lengths for these queries. The studies find evidence supporting the idea that different query types are best answered with summaries of different lengths.

2 Related Work

2.1 Query-biased Summaries

In the early days of the web, the result summary consisted of the first few lines of text, due both to concerns about intellectual property, and because often that was the only part of the full text that the search engines retained from their crawls. Eventually, search engines started showing what are known variously as *query-biased summaries*, *keyword-in-*

context (KWIC) extractions, and *user-directed summaries* (Tombros and Sanderson, 1998). In these summaries, sentence fragments, full sentences, or groups of sentences that contain query terms are extracted from the full text. Early versions of this idea were developed in the Snippet Search tool (Pedersen et al., 1991) and the Superbook tool's Table-of-Contents view (Egan et al., 1989).

A query-biased summary shows sentences that summarize the ways the query terms are used within the document. In addition to showing which subsets of query terms occur in a retrieved document, this display also exposes the context in which the query terms appear with respect to one another.

Research suggests that query-biased summaries are superior to showing the first few sentences from documents. Tombros & Sanderson (1998), in a study with 20 participants using TREC *ad hoc* data, found higher precision and recall and higher subjective preferences for query-biased summaries over summaries showing the first few sentences. Similar results for timing and subjective measurements were found by White et al. (2003) in a study with 24 participants. White et al. (2003) also describe experiments with different sentence selection mechanisms, including giving more weight to sentences that contained query words along with text formatting.

There are significant design questions surrounding how best to formulate and display query-biased summaries. As with standard document summarization and extraction, there is an inherent trade-off between showing long, informative summaries and minimizing the screen space required by each search hit. There is also a tension between showing short snippets that contain all or most of the query terms and showing coherent stretches of text. If the query terms do not co-occur near one another, then the extract has to become very long if full sentences and all query terms are to be shown. Many web search engine snippets compromise by showing fragments instead of sentences.

2.2 Studies Comparing Results Lengths

Recently, a few studies have analyzed the results of varying search summary length.

In the question-answering context (as opposed to general web search), Lin et al. (2003) conducted a usability study with 32 computer science students comparing four types of answer context: exact answer, answer-in-sentence, answer-in-paragraph, and answer-in-document. To remove effects of incorrect answers, they used a system that produced only correct answers, drawn from an online encyclopedia. Participants viewed answers for 8 question scenarios. Lin et al. (2003) found no significant differences in task completion times, but they did find differences in subjective responses. Most participants (53%) preferred paragraph-sized chunks, noting that a sentence wasn't much more information beyond the exact answer, and a full document was oftentimes too long. That said, 23% preferred full documents, 20% preferred sentences, and one participant preferred exact answer, thus suggesting that there is considerable individual variation.

Paek et al. (2004) experimented with showing differing amounts of summary information in results listings, controlling the study design so that only one result in each list of 10 was relevant. For half the test questions, the target information was visible in the original snippet, and for the other half, the participant needed to use their mouse to view more information from the relevant search result. They compared three interface conditions:

(i) a standard search results listing, in which a mouse click on the title brings up the full text of the web page,

(ii) "instant" view, for which a mouseclick expanded the document summary to show additional sentences from the document, and those sentences contained query terms and the answer to the search task, and

(iii) a "dynamic" view that responded to a mouse hover, and dynamically expanded the summary with a few words at a time.

Eleven out of 18 participants preferred instant view over the other two views, and on average all participants produced faster and more accurate results with this view. Seven participants preferred dynamic view over the others, but many others found this view disruptive. The dynamic view suffered from the problem that, as the text expanded, the mouse no longer covered the selected results, and so an unintended, different search result sometimes started to expand. Notably, none of the participants preferred the standard results listing view.

Cutrell & Guan (2007), compared search summaries of varying length: short (1 line of text), medium (2-3 lines) and long (6-7 lines) using search engine-produced snippets (it is unclear if the summary text was contiguous or included ellipses). They also compared 6 navigational queries (where the goal is to find a website's homepage), with 6 informational queries (e.g., "find when the Titanic set sail for its only voyage and what port it left from," "find out how long the Las Vegas monorail is"). In a study with 22 participants, they found that participants were 24 seconds faster on average with the long view than with the short and medium view. The also found that participants were 10 seconds slower on average with the long view for the navigational tasks. They present eye tracking evidence which suggests that on the navigational task, the extra text distracts the eye from the URL. They did not report on subjective responses to the different answer lengths.

Rose et al. (2007) varied search results summaries along several dimensions, finding that text choppiness and sentence truncation had negative effects, and genre cues had positive effects. They did not find effects for varying summary length, but they only compared relatively similar summary lengths (2 vs. 3 vs. 4 lines long).

2.3 Categorizing Questions by Expected Answer Types

In the field of automated question-answering, much effort has been expended on automatically determining the kind of answer that is expected for a given question. The candidate answer types are often drawn from the types of questions that have appeared in the TREC Question Answering track (Voorhees, 2003). For example, the Webclopedia project created a taxonomy of 180 types of question targets (Hovy et al., 2002), and the FALCON project (Harabagiu et al., 2003) developed an answer taxonomy with 33 top level categories (such as PERSON, TIME, REASON, PRODUCT, LOCATION, NUMERICAL VALUE, QUOTATION), and these were further refined into an unspecified number of additional categories. Ramakrishnan et al. (2004) show an automated method for determining expected answer types using syntactic information and mapping query terms to WordNet.

2.4 Categorizing Web Queries

A different line of research is the query log categorization problem. In query logs, the queries are often much more terse and ill-defined than in the TREC QA track, and, accordingly, the taxonomies used to classify what is called the query intent have been much more general.

In an attempt to demonstrate how information needs for web search differ from the assumptions of pre-web information retrieval systems, Broder (2002) created a taxonomy of web search goals, and then estimated frequency of such goals by a combination of an online survey (3,200 responses, 10% response rate) and a manual analysis of 1,000 query from the AltaVista query logs. This taxonomy has been heavily influential in discussions of query types on the Web.

Rose & Levinson (2004) followed up on Broder's work, again using web query logs, but developing a taxonomy that differed somewhat from Broder's. They manually classified a set of 1,500 AltaVista search engine log queries. For two sets of 500 queries, the labeler saw just the query and the retrieved documents; for the third set the labeler also saw information about which item(s) the searcher clicked on. They found that the classifications that used the extra information about clickthrough did not change the proportions of assignments to each category. Because they did not directly compare judgments with and without click information on the same queries, this is only weak evidence that query plus retrieved documents is sufficient to classify query intent.

Alternatively, queries from web query logs can be classified according to the *topic* of the query, independent of the type of information need. For example, a search involving the topic of weather can consist of the simple information need of looking at today's forecast, or the rich and complex information need of studying meteorology. Over many years, Spink & Jansen et al. (2006; 2007) have manually analyzed samples of query logs to track a number of different trends. One of the most notable is the change in topic mix. As an alternative to manual classification of query topics, Shen et al. (2005) described an algorithm for automatically classifying web queries into a set of pre-defined topics. More recently, Broder et al. (2007) presented a highly accurate method (around .7 F-score) for classifying short, rare queries into a taxonomy of 6,000 categories.

3 Study Goals

Related work suggests that longer results are preferable, but not for all query types. The goal of our efforts was to determine preferred result length for search results, depending on type of query. To do this, we performed two studies:

1. We asked a set of judges to categorize a large set of web queries according to their expected preferred response type and expected preferred response length.

2. We then developed high-quality answer passages of different lengths for a subset of these queries by selecting appropriate passages from the online encyclopedia Wikipedia, and asked judges to rate the quality of these answers.

The results of this study should inform search interface designers about what the best presentation format is.

3.1 Using Mechanical Turk

For these studies, we make use of a web service offered by Amazon.com called Mechanical Turk, in which participants (called "turkers") are paid small sums of money in exchange for work on "Human Intelligence tasks" (HITs).[1] These HITs are generated from an XML description of the task created by the investigator (called a "requester"). The participants can come from any walk of life, and their identity is not known to the requesters. We have in past work found the results produced by these judges to be of high quality, and have put into place various checks to detect fraudulent behavior. Other researchers have investigated the efficacy of language annotation using this service and have found that the results are of high quality (Su et al., 2007).

1. Person(s)
2. Organization(s)
3. Time(s) (date, year, time span etc.)
4. Number or Quantity
5. Geographic Location(s) (e.g., city, lake, address)
6. Place(s) (e.g.,”the White House”, “at a supermarket”)
7. Obtain resource online (e.g., movies, lyrics, books, magazines, knitting patterns)
8. Website or URL
9. Purchase and product information
10. Gossip and celebrity information
11. Language-related (e.g., translations, definitions, crossword puzzle answers)
12. General information about a topic
13. Advice
14. Reason or Cause, Explanation
15. Yes/No, with or without explanation or evidence
16. Other
17. Unjudgable

Table 1: Allowable responses to the question: "What sort of result or results does the query ask for?" in the first experiment.

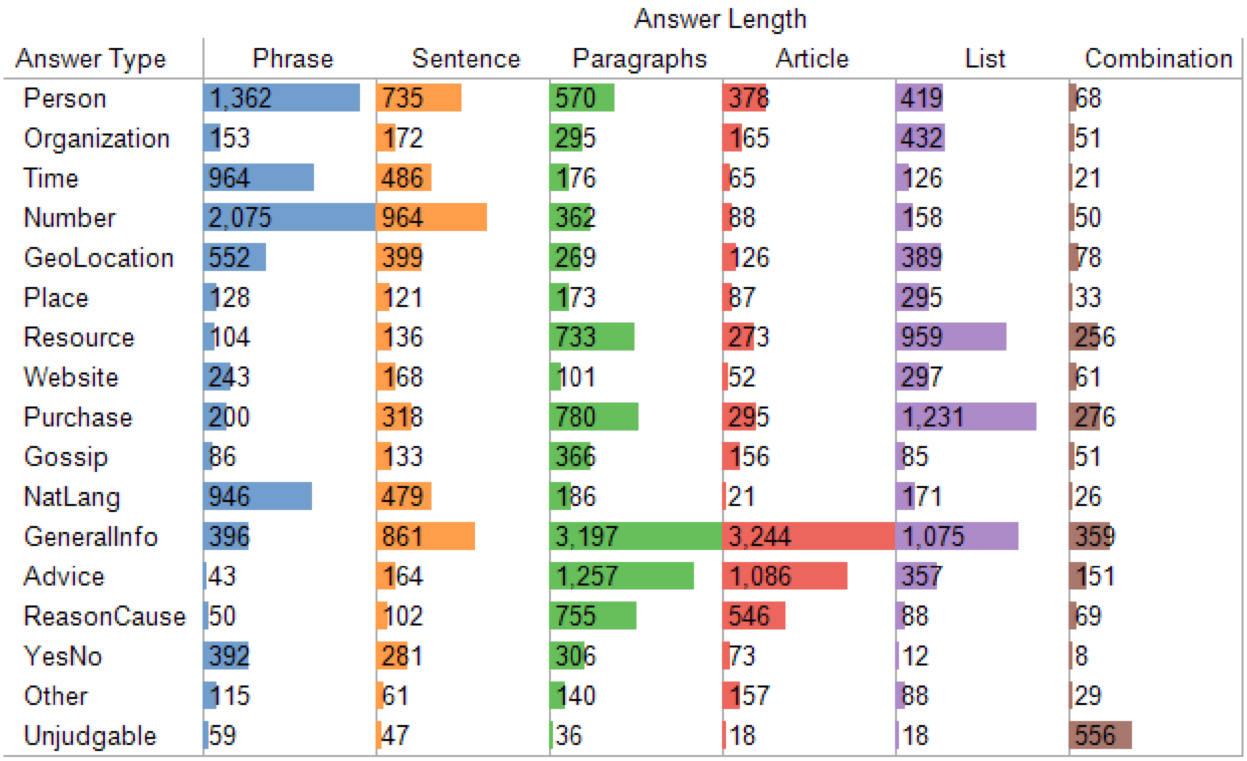
1. A word or short phrase
2. A sentence
3. One or more paragraphs (i.e. at least several sentences)
4. An article or full document
5. A list
6. Other, or some combination of the above

Table 2: Allowable responses to the question: "How long is the best result for this query?" in the first experiment.

3.2 Estimating Ideal Answer Length and Type

We developed a set of 12,790 queries, drawn from Powerset's in house query database which contains representative subsets of queries from different search engines' query logs, as well as hand-edited query sets used for regression testing. There are a disproportionally large number of natural language queries in this set compared with query sets from typical keyword engines. Such queries are often complete questions and are sometimes grammatical fragments (e.g., "date of next US election") and so are likely to be amenable to interesting natural language processing algorithms, which is an area of in-

[1] Website: http://www.mturk.com. For experiment 1, approximately 38,000 HITs were completed at a cost of about $1,500. For experiment 2, approximately 7,300 HITs were completed for about $170. Turkers were paid between $.01 and $.05 per HIT depending on task complexity; Amazon imposes additional charges.



Answer Type	Answer Length Phrase	Sentence	Paragraphs	Article	List	Combination
Person	1,362	735	570	378	419	68
Organization	153	172	295	165	432	51
Time	964	486	176	65	126	21
Number	2,075	964	362	88	158	50
GeoLocation	552	399	269	126	389	78
Place	128	121	173	87	295	33
Resource	104	136	733	273	959	256
Website	243	168	101	52	297	61
Purchase	200	318	780	295	1,231	276
Gossip	86	133	366	156	85	51
NatLang	946	479	186	21	171	26
GeneralInfo	396	861	3,197	3,244	1,075	359
Advice	43	164	1,257	1,086	357	151
ReasonCause	50	102	755	546	88	69
YesNo	392	281	306	73	12	8
Other	115	61	140	157	88	29
Unjudgable	59	47	36	18	18	556

Figure 1: Results of the first experiment. The y-axis shows the semantic type of the predicted answer, in the same order as listed in Table 1; the x-axis shows the preferred length as listed in Table 2. Three bars with length greater than 1,500 are trimmed to the maximum size to improve readability (GeneralInfo/Paragraphs, GeneralInfo/Article, and Number/Phrase).

terest of our research. The average number of words per query (as determined by white space separation) was 5.8 (sd. 2.9) and the average number of characters (including punctuation and white space) was 32.3 (14.9). This is substantially longer than the current average for web search query, which was approximately 2.8 in 2005 (Jansen et al., 2007); this is due to the existence of natural language queries.

Judges were asked to classify each query according to its expected response type into one of 17 categories (see Table 1). These categories include answer types used in question answering research as well as (to better capture the diverse nature of web queries) several more general response types such as *Advice* and *General Information*. Additionally, we asked judges to anticipate what the best result length would be for the query, as shown in Table 2.

Each of the 12,790 queries received three assessments by MTurk judges. For answer *types*, the number of times all three judges agreed was 4537 (35.4%); two agreed 6030 times (47.1%), and none agreed 2223 times (17.4%). Not surprisingly, there was significant overlap between the label *GeneralInfo* and the other categories. For answer *length* estimations, all three judges agreed in 2361 cases (18.5%), two agreed in 7210 cases (56.4%) and none 3219 times (25.2%).

Figure 1 summarizes expected length judgments by estimated answer category. Distribution of the length categories differs a great deal across the individual expected response categories. In general, the results are intuitive: judges preferred short responses for "precise queries" (e.g., those asking for numbers) and they preferred longer responses for queries in broad categories like *Advice* or *GeneralInfo*. But some results are less intuitive: for example, judges preferred different response lengths for queries categorized as *Person* and *Organization* – in fact for the latter the largest single selection made was *List*. Reviewing the queries for these two categories, we note that most queries about organizations in our collection asked for companies

length type	average	std dev
Word or Phrase	38.1	25.8
Sentence	148.1	71.4
Paragraph	490.5	303.1
Section	1574.2	1251.1

Table 3: Average number of characters for each answer length type for the stimuli used in the second experiment.

(e.g. "around the world travel agency") and for these there usually is more than one correct answer, whereas the queries about persons ("CEO of microsoft") typically only had one relevant answer. The results of this table show that there are some trends but not definitive relationships between query type (as classified in this study) and expected answer length. More detailed classifications might help resolve some of the conflicts.

3.3 Result Length Study

The purpose of the second study was twofold: first, to see if doing a larger study confirms what is hinted at in the literature: that search result lengths longer than the standard snippet may be desirable for at least a subset of queries. Second, we wanted to see if judges' predictions of desirable results lengths would be confirmed by other judges' responses to search results of different lengths.

3.3.1 Method

It has been found that obtaining judges' agreement on intent of a query from a log can be difficult (Rose and Levinson, 2004; Kellar et al., 2007). In order to make the task of judging query relevance easier, for the next phase of the study we focused on only those queries for which all three assessors in the first experiment agreed both on the category label and on the estimated ideal length. There were 1099 such high-confidence queries, whose average number of words was 6.3 (2.9) and average number of characters was 34.5 (14.3).

We randomly selected a subset of the high-agreement queries from the first experiment and manually excluded queries for which it seemed obvious that no responses could be found in Wikipedia. These included queries about song lyrics, since intellectual property restrictions prevent these being posted, and crossword puzzle questions such as "a four letter word for water."

The remaining set contained 170 queries. MTurk annotators were asked to find one good text passage (in English) for each query from the Wikipedia online encyclopedia. They were also asked to subdivide the text of this answer into each of the following lengths: a word or phrase, a sentence, a paragraph, a section or an entire article.[2] Thus, the shorter answer passages are subsumed by the longer ones.

Table 3 shows the average lengths and standard deviations of each result length type. Table 4 contains sample answers for the shorter length formats for one query. For 24 of the 170 queries the annotators could not find a suitable response in Wikipedia, e.g., "How many miles between NY and Milwaukee?" We collected two to five results for each of the remaining 146 queries and manually chose the best of these answer passages. Note that, by design, all responses were factually correct; they only differed in their length.

Ten MTurk judges saw each query/answer length pair, and for each of these, were told: "Below you see a search engine query and a possible response. We would like you to give us your opinion about the response. We are especially interested in the length of the response. Is it suitable for the query? Is there too much or not enough information? Please rate the response on a scale from 0 (very bad response) to 10 (very good response)." There were 124 judges in total; of these, 16 did more than 146 HITs, meaning they saw the same query more than one time (but with different lengths). Upon examination of the results, we determined that two of these high-volume judges were not trying to do the task properly, and so we dropped their judgments from the final analysis.

3.3.2 Results

Our results show that judges prefer results of different lengths, depending on the query. The results also suggest that judges' estimates of a preferred result length in the first experiment are accurate predictors when there is strong agreement among them. Figure 2 shows in four diagrams

[2] Note the slight difference between the length categories in the first and second experiment: The ***List*** and ***Other*** options were dropped for the second experiment because we wanted to concentrate on textual length. Additionally, to provide more than one option between ***Sentence*** and ***Article***, the category ***One or more paragraphs*** was split up into two: ***(One) Paragraph*** and ***(One) Section***.

query	Who was the first person to scale K2?
Paragraph	An Italian expedition finally succeeded in ascending to the summit of K2 on July 31, 1954. The expedition was led by Ardito Desio, although the two climbers who actually reached the top were Lino Lacedelli and Achille Compagnoni. The team included a Pakistani member, Colonel Muhammad Ata-ullah. He had been a part of an earlier 1953 American expedition which failed to make the summit because of a storm which killed a key climber, Art Gilkey. On the expedition also was the famous Italian climber Walter Bonatti. He proved vital to the expeditions success in that he carried vital oxygen to 26,600ft for Lacedelli and Compagnoni. His dramatic bivouac, at that altitude with the equipment, wrote another chapter in the saga of Himalayan climbing.
Sentence	The expedition was led by Ardito Desio, although the two climbers who actually reached the top were Lino Lacedelli and Achille Compagnoni.
Phrase	Lino Lacedelli and Achille Compagnoni

Table 4: Sample answers of differing lengths used as input for the second study. Note that the shorter answers are contained in the longer ones. For the full article case, judges were asked to follow a hyperlink to an article.

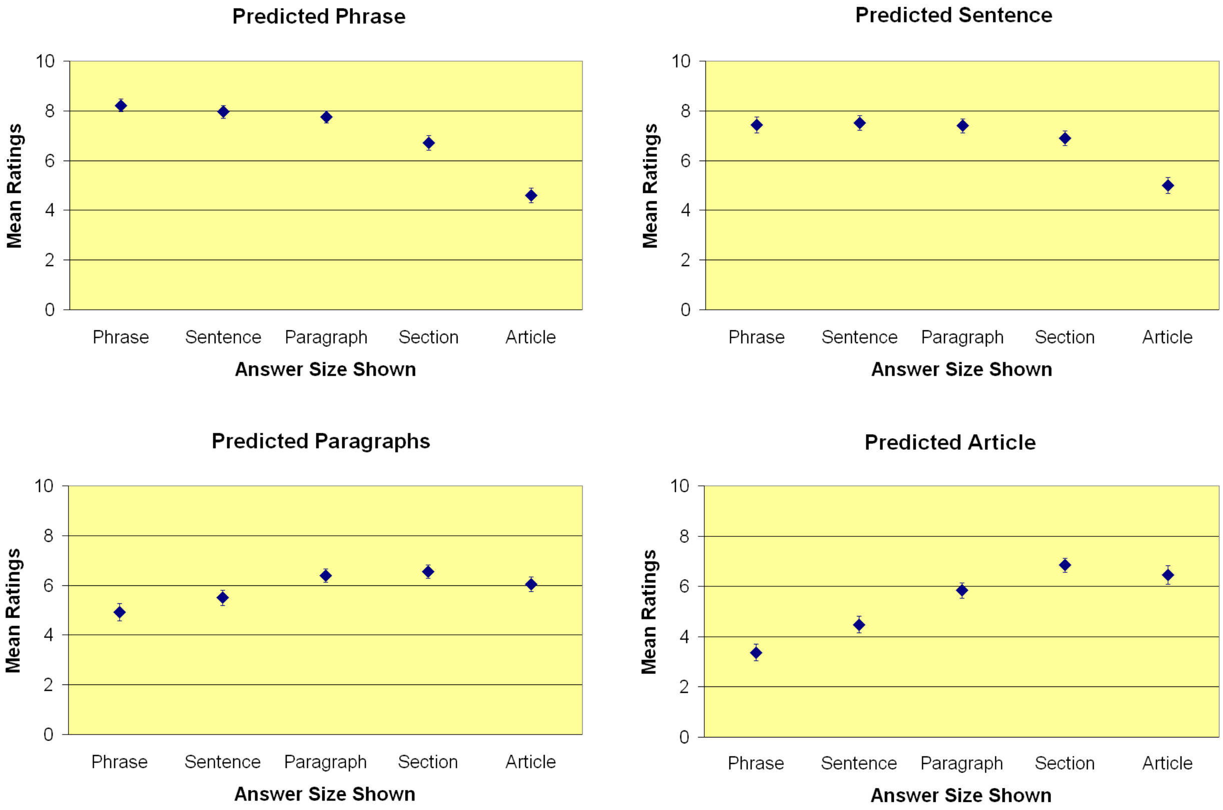


Figure 2: Results of the second experiment, where each query/answer-length pair was assessed by 8–10 judges using a scale of 0 (‘very bad’) to 10 (‘very good’). Marks indicate means and standard errors. The top left graph shows responses of different lengths for queries that were classified as *best answered with a phrase* in the first experiment. The upper right shows responses for queries predicted to be *best answered with a sentence*, lower left for *best answered with one or more paragraphs* and lower right for *best answered with an article*.

	Slope	Std. Error	p-value
Phrase	-0.850	0.044	$<$ 0.0001
Sentence	-0.550	0.050	$<$ 0.0001
Paragraph	0.328	0.049	$<$ 0.0001
Article	0.856	0.053	$<$ 0.0001

Table 5: Results of unweighted linear regression on the data for the second experiment, which was separated into four groups based on the predicted preferred length.

how queries assigned by judges to one of the four length categories from the first experiment were judged when presented with responses of the five answer lengths from the second experiment. The graphs show the means and standard error of the judges' scores across all queries for each predicted-length/presented-length combination.

In order to test whether these results are significant we performed four separate linear regressions; one for each of the predicted preferred length categories. The snippet length, the independent variable, was coded as 1-5, shortest to longest. The score for each query-snippet pair is the dependent variable. Table 5 shows that for each group there is evidence to reject the null hypothesis that the slope is equal to 0 at the 99% confidence level. High scores are associated with shorter snippet lengths for queries with predicted preferred length *phrase* or *sentence* and also with longer snippet lengths for queries with predicted preferred length *paragraphs* or *article*. These associations are strongest for the queries with the most extreme predicted preferred lengths (*phrase* and *article*).

Our results also suggest the intuition that the best answer lengths do not form strictly distinct classes, but rather lie on a continuum. If the ideal response is from a certain category (e.g., a sentence), returning a result from an adjacent category (a phrase or a paragraph) is not strongly penalized by judges, whereas retuning a result from a category further up or down the scale (an article) is.

One potential drawback of this study format is that we do not show judges a list of results for queries, as is standard in search engines, and so they do not experience the tradeoff effect of longer results requiring more scrolling if the desired answer is not shown first. However, the earlier results of Cutrell & Guan (2007) and Paek et al. (2004) suggest that the preference for longer results occurs even in contexts that require looking through multiple results. Another potential drawback of the study is that judges only view one relevant result; the effects of showing a list of long non-relevant results may be more negative than that of showing short non-relevant results; this study would not capture that effect.

4 Conclusions and Future Work

Our studies suggest that different queries are best served with different response lengths (Experiment 1), and that for a subset of especially clear queries, human judges can predict the preferred result lengths (Experiment 2). The results furthermore support the contention that standard results listings are too short in many cases, at least assuming that the summary shows information that is relevant for the query. These findings have important implications for the design of search results presentations, suggesting that as user queries become more expressive, search engine results should become more responsive to the type of answer desired. This may mean showing more context in the results listing, or perhaps using more dynamic tools such as expand-on-mouseover to help answer the query in place.

The obvious next step is to determine how to automatically classify queries according to their predicted result length and type. For classifying according to expected length, we have run some initial experiments based on unigram word counts which correctly classified 78% of 286 test queries (on 805 training queries) into one of three length bins. We plan to pursue this further in future work. For classifying according to type, as discussed above, most automated query classification for web logs have been based on the topic of the query rather than on the intended result type, but the question answering literature has intensively investigated how to predict appropriate answer types. It is likely that the techniques from these two fields can be productively combined to address this challenge.

Acknowledgments. This work was supported in part by Powerset, Inc., and in part by Microsoft Research through the MSR European PhD Scholarship Programme. We would like to thank Susan Gruber and Bonnie Webber for their helpful comments and suggestions.

References

A. Broder, M. Fontoura, E. Gabrilovich, A. Joshi, V. Josifovski, and T. Zhang. 2007. Robust classification of rare queries using web knowledge. *Proceedings of SIGIR 2007*.

A. Broder. 2002. A taxonomy of web search. *ACM SIGIR Forum*, 36(2):3–10.

E. Cutrell and Z. Guan. 2007. What Are You Looking For? An Eye-tracking Study of Information Usage in Web Search. *Proceedings of ACM SIGCHI 2007*.

D.E. Egan, J.R. Remde, L.M. Gomez, T.K. Landauer, J. Eberhardt, and C.C. Lochbaum. 1989. Formative design evaluation of Superbook. *ACM Transactions on Information Systems (TOIS)*, 7(1):30–57.

S.M. Harabagiu, S.J. Maiorano, and M.A. Pasca. 2003. Open-domain textual question answering techniques. *Natural Language Engineering*, 9(03):231–267.

E. Hovy, U. Hermjakob, and D. Ravichandran. 2002. A question/answer typology with surface text patterns. *Proceedings of the second international conference on Human Language Technology Research*, pages 247–251.

B.J. Jansen and Spink. 2006. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing and Management*, 42(1):248–263.

B.J. Jansen, A. Spink, and S. Koshman. 2007. Web searcher interaction with the Dogpile.com metasearch engine. *Journal of the American Society for Information Science and Technology*, 58(5):744–755.

M. Kellar, C. Watters, and M. Shepherd. 2007. A Goal-based Classification of Web Information Tasks. *JASIST*, 43(1).

J. Lin, D. Quan, V. Sinha, K. Bakshi, D. Huynh, B. Katz, and D.R. Karger. 2003. What Makes a Good Answer? The Role of Context in Question Answering. *Human-Computer Interaction (INTERACT 2003)*.

G. Marchionini, R.W. White, and Marchionini. 2008. Find What You Need, Understand What You Find. *Journal of Human-Computer Interaction (to appear)*.

T. Paek, S.T. Dumais, and R. Logan. 2004. WaveLens: A new view onto internet search results. *Proceedings on the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 727–734.

J. Pedersen, D. Cutting, and J. Tukey. 1991. Snippet search: A single phrase approach to text access. *Proceedings of the 1991 Joint Statistical Meetings*.

G. Ramakrishnan and D. Paranjpe. 2004. Is question answering an acquired skill? *Proceedings of the 13th international conference on World Wide Web*, pages 111–120.

D.E. Rose and D. Levinson. 2004. Understanding user goals in web search. *Proceedings of the 13th international conference on World Wide Web*, pages 13–19.

D.E. Rose, D. Orr, and R.G.P. Kantamneni. 2007. Summary attributes and perceived search quality. *Proceedings of the 16th international conference on World Wide Web*, pages 1201–1202.

D. Shen, R. Pan, J.T. Sun, J.J. Pan, K. Wu, J. Yin, and Q. Yang. 2005. Q2C@UST: our winning solution to query classification in KDDCUP 2005. *ACM SIGKDD Explorations Newsletter*, 7(2):100–110.

Q. Su, D. Pavlov, J. Chow, and W. Baker. 2007. Internet-Scale Collection of Human-Reviewed Data. *Proceedings of WWW 2007*.

A. Tombros and M. Sanderson. 1998. Advantages of query biased summaries in information retrieval. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 2–10.

E.M. Voorhees. 2003. Overview of the TREC 2003 Question Answering Track. *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*.

R.W. White, J. Jose, and I. Ruthven. 2003. A task-oriented study on the influencing effects of query-biased summarisation in web searching. *Information Processing and Management*, 39(5):707–733.