

# Analysis of Factoid Questions for Effective Relation Extraction

Eugene Agichtein, Silviu Cucerzan, and Eric Brill

Microsoft Research,

One Microsoft Way, Redmond, WA, USA

{eugeneag, silviu, brill}@microsoft.com

## ABSTRACT

We present an analysis of the structured relationships observed in a randomly sampled set of question-like queries submitted to a search engine for a popular online encyclopedic document collection. Our study shows that a relatively small number of binary relationships account for most of the queries in the sample. This empirically validates an approach of analyzing query logs to identify the relationships most relevant to user needs and populating corresponding fact tables from the collection for factoid question answering. Our analysis shows that such an approach can lead to substantial coverage of user questions.

## Categories and Subject Descriptors

**H.3.1 [INFORMATION STORAGE AND RETRIEVAL]: Information Search and Retrieval; Question-answering**

**General Terms** Experimentation, Evaluation

**Keywords** Q&A, relation extraction, query log analysis.

## 1. INTRODUCTION AND PRIOR WORK

Coverage of questions asked and accuracy of answers returned are crucial aspects of automatic question answering (QA). Both issues can be addressed in a user-need-guided framework of extracting and storing *relevant* facts from large document collections.

Factoid questions comprise a significant fraction of all user queries submitted to search engines [10]. There has been a rising interest in extracting structured tables of facts and patterns to be used in answering factoid questions quickly and accurately [5]. Furthermore, gathering collections of reliable fact tables is an important step towards answering complex factoid questions [3, 7] by decomposing them into simpler questions [8] that can be answered using such tables.

Many existing QA systems (e.g., [1, 3, 6]) are optimized for the TREC competition, and may suffer from low coverage when used for a broader range of user questions. This difference may be further amplified when a QA system is used over a collection of documents that is dramatically different from the TREC document collections. By focusing on the relations needed to answer users' questions over a collection, we can substantially increase the effective coverage (i.e., fraction of queries that can be answered).

Our contributions include:

- Principled analysis of the structured relation space required to answer actual questions;
- Empirical confirmation of Zipf-like distribution and temporal stability of the types of user factoid questions;
- Empirical confirmation that previously studied relations are important to answering questions submitted by real users, but also provide a relatively low coverage of *question instances* (i.e., tokens) to an encyclopedic collection such as Encarta;
- Construction of a new annotated resource for IR and QA (Available at <http://research.microsoft.com/~eugeneag/sigir05/>).

## 2. DATASET AND ANNOTATION

We now describe our experimental methodology for exploring the space of relations required to answer user questions *to an actual document collection*. We used an Encarta query log, from which *singletons* were removed. This query log contains several million *query types* accounting for the queries with at least two *instances* submitted to the Encarta search engine in 2003 and 2004.

We were interested both in the overall composition of the question-like queries in the log and in identifying *factoid questions* that can be answered using relations extracted from the target document collection. We define factoid questions as questions that have a short answer (typically a noun phrase or a simple verb phrase) or an enumeration of such short answers. Among those, we were particularly interested in *binary factoid questions* (BF), i.e., correspond to a binary relation (e.g., “how tall is mount Everest?” → *height* (mount Everest, 29,035 ft)) without complex dependencies (e.g., “how old was Leonardo da Vinci when he painted Mona Lisa?” requires such a dependency). We started with a set of 9 predefined entity types (*person*, *location*, *organization*, *date*, *event*, *quantity*, *object*, *concept*, *animate*) and 12 relationships between these entities.

To focus our annotation efforts on likely factoid questions, we selected from our query logs a set of 256,508 question-like query types beginning with *who*, *when*, *what*, *where*, and *how*. A pool of queries was then selected at random from this set for annotation, proportionally to the number of instances of each query type in the log. Several annotators (with library, indexing, and computer science backgrounds) manually annotated a total of 2,215 queries using an interface that captured each of their action (searching Encarta, searching the web, browsing documents, etc.). For each query, the annotators were asked to do the following:

- Decide whether a query is a BF question, an *Explanation* (a question that requires a multiple-sentence answer), a *Definition* (a question that can be answered by dictionary lookup), a *Navigational* query, or *Other* (if the query intent was not clear);
- For BF questions, the annotators were asked to choose a binary relationship from a predefined list or, if none matched, to create a new relationship of one of the following two types: *has property* (e.g. “how long was Columbus' journey” → *<event> has duration <quantity>*) and *performs action* (e.g. “what pandas eat” → *<animate> eats <object>*).

Annotators added two new entity types (*language* and *title*) to our initial list, and created a total of 144 relations (available at the website mentioned above).

Type	Query count	Instance count	Example
<b>BF Question</b>	<b>704</b>	<b>10193</b>	<b>who invented the car</b>
Explanation	687	9340	how to build a house
Definition	399	8087	what is mild steel
Navigational	14	4038	how do I find yahoo?
Other	411	30	how to react on moods
<i>Total:</i>	<b>2215</b>	<b>31688</b>	

Table 1: Statistics on query types in the sample

Table 1 summarizes the structure of our query sample analyzed in the given timeframe (30 hours). Approximately one third of the queries (704 of 2,215) were labeled as BF questions.

Annotators encountered two types of problems in classifying queries. One was deciding whether a query can be modeled as a BF question. When at least one annotator assigned a query to BF, the others also assigned it to this category 68% of the time, the most frequent conflicting assignment being *explanation* (e.g. “what caused the American revolution” → *explanation* vs. *<event> has cause <concept>*). Once the annotators agreed on classifying a query as BF, they also agreed 57% of the time on the specific relation to assign to it. This lower than expected agreement was due to a second problem: the relations chosen were often semantically the same, but differently named (e.g. “who made dynamite” → *<person> invented <object>*, *<person> discovers <object>*); After conflating identical-meaning relations, we observed a pairwise annotator agreement between 75% to 94% on the post-processed data, with an average agreement of 84%.

### 3. RESULTS

Figure 1 reports the percentage of relations needed to cover the BF questions in the sample (reported as “Exact”). We also report the coverage when the relationships are identified only by their verb (e.g., *invented*), and ignoring the entity types.

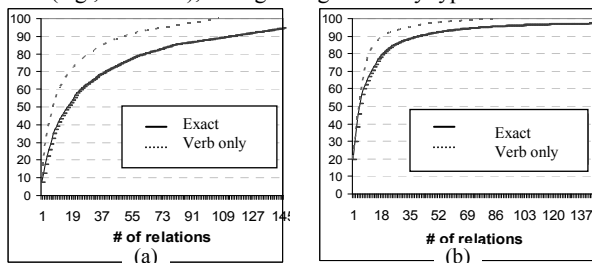


Figure 1: Number of unique relations to cover % of annotated factoid query types (a) and query instances (b) (in the sample).

Relation	Type %	Instance %	Literature
<person> discovers <concept>	7.7	2.9	[9]
<person> has position <concept>	5.6	4.6	[4][5][9]
<location> has location <location>	5.2	1.5	[9]
<person> known for <concept>	4.7	1.7	[9]
<event> has date <date>	4.1	0.9	-
<object> has discovery date <date>	3.3	1.0	-
<person> creates <object>	3.3	1.5	-
<animate> eats <object>	2.9	1.8	-
<event> has location <location>	2.4	1.6	-
<object> has alias <title>	2.3	0.7	-
<i>Total coverage</i>	41.5	18.2	-

Table 2: Top 10 most frequent relations (by query type)

Relation	Type %	Instance %	Literature
<location> has neighbors <location>	0.5	21.2	-
<location> has founding date <date>	1.4	11.0	-
<animate> has speed <quantity>	0.2	9.0	-
<animate> has color <color>	0.3	8.0	-
<location> has length <quantity>	0.2	6.7	-
<i>Total coverage</i>	2.6	55.9	-

Table 3: Top 5 most frequent relations (by query instance)

Table 2 reports the 10 most frequent relations observed, accounting for more than 41% of the query types in our sample.

Furthermore, Table 3 shows that as few as 5 relations cover more than 55% of the actual question instances in the log. Interestingly, these relations are disjoint from those in Table 2.

We now investigate the temporal stability of the set of relations required to answer user questions. We measure how many new relationships are needed to cover the queries observed at a given point in time given the set of relationships created before that time. To simulate this, we split the time axis into one-month intervals and assign to each interval the relationships created for queries in the labeled sample that had realizations in that interval. As shown in Figure 4.2, after 4 months, the relationships collected from the sample “previously” cover at least 80% of the queries at each point in time. This confirms our hypothesis that the most productive relations remain consistently interesting for the users.

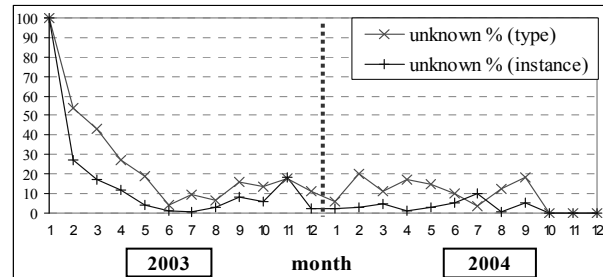


Figure 2: Relation coverage of future factoid queries

### 4. CONCLUSIONS

We consider this work a promising step towards answering factoid questions using fact tables extracted from trusted resources. The skewed distribution of relationships observed in the annotated queries indicates that a limited number of fact tables can cover the bulk of user factoid questions. This approach could be extended to handling complex questions through decomposition into binary factoid questions.

Furthermore, we advocate focusing computational and annotation resources on extracting fact tables for frequently queried relationships, and on mapping user questions to appropriate relations. We envision a framework where logs of queries to trusted resources or the web are periodically analyzed, suggesting the most beneficial updates to collections of extracted fact tables.

### 5. REFERENCES

- [1] E. Brill, S. Dumais, M. Banko. An Analysis of the AskMSR Question-Answering System, *EMNLP 2002*
- [2] A. Broder, Taxonomy of Web Search, *SIGIR Forum*, 2002
- [3] J. Chu-Carroll et al. IBM’s PIQUANT II in TREC 2004, *TREC 2004*
- [4] Etzioni et al., Web-scale information extraction in Knowitall: preliminary results, *WWW 2004*
- [5] M. Fleischman, E. Hovy, A. Echiabi, Offline Strategies for Online Question Answering: Answering Questions Before They are Asked, *ACL 2003*
- [6] E. Hovy, L. Gerber, U. Hermjakob, M. Junk, and C. Lin, Question Answering in Webclopedia. *TREC-9*, 2000
- [7] S. Harabagiu, D. Moldovan, P. Surdeanu, et al., Answering Complex, List and Context Questions with LCC’s QA Server, *TREC-10*, 2001
- [8] D. Moldovan, C. Clark, S. Harabagiu, S. Maiorano, COGEX: A Logic Prover for Question Answering, *ACL 2003*
- [9] D. Ravichandran and E. Hovy, Learning Surface Text Patterns for a Question Answering System, *ACL 2002*
- [10] Spink and H.C. Ozmutlu, Ask Jeeves query analysis: What do people ask?, *ASIST 2001*