Answering Definition Questions Using Multiple Knowledge Sources

Wesley Hildebrandt, Boris Katz, and Jimmy Lin
MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Cambridge, MA 02139
{wes,boris,jimmylin}@csail.mit.edu

Abstract

Definition questions represent a largely unexplored area of question answering—they are different from factoid questions in that the goal is to return as many relevant "nuggets" of information about a concept as possible. We describe a multi-strategy approach to answering such questions using a database constructed offline with surface patterns, a Web-based dictionary, and an off-the-shelf document retriever. Results are presented from component-level evaluation and from an end-to-end evaluation of our implemented system at the TREC 2003 Question Answering Track.

1 Introduction

To date, research in question answering has concentrated on factoid questions such as "Who was Abraham Lincoln married to?" The standard strategy for answering these questions using a textual corpus involves a combination of information retrieval and named-entity extraction technology; see (Voorhees, 2002) for an overview. Factoid questions, however, represent but one facet of question answering, whose broader goal is to provide humans with intuitive information access using natural language.

In contrast to factoid questions, the objective for "definition" questions is to produce as many useful "nuggets" of information as possible. For example, the answer to "Who is Aaron Copland?" might include the following:

American composer
wrote ballets and symphonies
born in Brooklyn, New York, in 1900
son of a Jewish immigrant
American communist
civil rights advocate

Until recently, definition questions remained a largely unexplored area of question answering. Standard factoid question answering technology, designed to extract single answers, cannot be directly applied to this task. The solution to this interesting research challenge will draw from related fields such as information extraction, multi-document summarization, and answer fusion.

In this paper, we present an approach to answering definition questions that combines knowledge from three sources. We present results from our own component analysis and the TREC 2003 Question Answering Track.

2 Answering Definition Questions

Our first step in answering a definition question is to extract the concept for which information is being sought—called the target term, or simply, the target. Once the target term has been found, three techniques are employed to retrieve relevant nuggets: lookup in a database created from the AQUAINT corpus[1], lookup in a Web dictionary followed by answer projection, and lookup directly in the AQUAINT corpus with an IR engine. Answers from the three different sources are then merged to produce the final system output. The following subsections describe each of these modules in greater detail.

2.1 Target Extraction

We have developed a simple pattern-based parser to extract the target term using regular expressions. If the natural language question does not fit any of our patterns, the parser heuristically extracts the last sequence of capitalized words in the question as the target.

Our simple target extractor was tested on all definition questions from the TREC-9 and TREC-10 QA Track testsets and performed with one hundred percent accuracy on those questions. However, there were several instances where the target term was not correctly extracted from

[1] official corpus used for the TREC QA Track, available from the Linguistic Data Consortium

Name		Pattern	Bindings
Copular[1]	(e1_is)	NP$_1$ be NP$_2$	[NP$_1$ = t, NP$_2$ = n]
Become[2]	(e1_beca)	NP$_1$ become NP$_2$	[NP$_1$ = t, NP$_2$ = n]
Verb[3]	(e1_verb):	NP$_1$ v NP$_2$	[where $v \in$ biography-verb; NP$_1$ = t, NP$_2$ = n]
Appositive[4]	(e1/2_appo)	NP$_1$, NP$_2$	[NP$_1$ = $t \vee n$, NP$_2$ = $t \vee n$]
Occupation[5]	(e2_occu)	NP$_1$ NP$_2$	[where head(NP$_1$) $\in$ occupation; NP$_1$ = n, NP$_2$ = t]
Parenthesis[6]	(e1_pare)	NP$_1$ (NP$_2$)	[NP$_1$ = t, NP$_2$ = n]
Also-known-as[7]	(e1/2_aka)	NP$_1$, (also) known as NP$_2$	[NP$_1$ = $t \vee n$, NP$_2$ = $t \vee n$]
Also-called[8]	(e2_also)	NP$_1$, (also) called NP$_2$	[NP$_1$ = n, NP$_2$ = t]
Or[9]	(e1_or)	NP$_1$, or NP$_2$	[NP$_1$ = t, NP$_2$ = n]
Like[10]	(e2_like)	NP$_1$ (such as$\vert$like) NP$_2$	[NP$_1$ = n, NP$_2$ = t]
Relative clause[11]	(e1_wdt)	NP (which$\vert$that) VP	[NP = t, VP = n]

[1] In order to filter out spurious nuggets (e.g., progressive tense), our system discards nuggets that do not begin with a determiner.
[2] The verb *become*, like *be*, often yields good nuggets that define a target.
[3] By statistically analyzing a corpus of biographies of famous people, we compiled a list of verbs commonly used to describe people and their accomplishments, such as *write*, *invent*, and *make*.
[4] Either NP$_1$ or NP$_2$ can be the target; thus, we index both NPs as the target term.
[5] NPs preceding proper nouns provide information such as occupation or affiliation. To boost precision, our system discards nuggets that do not contain an occupation (e.g., *actor*, *spokesman*, *leader*). We mined this list from WordNet and the Web.
[6] Parenthetical expressions usually contain interesting nuggets; for persons, they often include a lifespan or job description.
[7] Either NP$_1$ or NP$_2$ can be the target; thus, we index both NPs as the target term.
[8] This and the previous pattern frequently identify hyponymy relations; typically, NP$_1$ is the hypernym of NP$_2$.
[9] This pattern often identifies the discourse function of elaboration.
[10] This pattern typically identifies an exemplification relationship, where NP$_2$ is an instance of NP$_1$.
[11] Relative clauses often provide useful nuggets.

Table 1: Description of the surface patterns used in constructing our database. (t is short for target, n for nugget)

the definition questions in TREC 2003, which made it difficult for downstream modules to find relevant nuggets (see Section 3.2 for a discussion).

2.2 Database Lookup

The use of surface patterns for answer extraction has proven to be an effective strategy for factoid question answering (Soubbotin and Soubbotin, 2001; Brill et al., 2001; Hermjakob et al., 2002). Typically, surface patterns are applied to a candidate set of documents returned by a document or passage retriever. Although this strategy often suffers from low recall, it is generally not a problem for factoid questions, where only a single instance of the answer is required. Definition questions, however, require a system to find as many relevant nuggets as possible, making recall very important.

To boost recall, we employed an alternative strategy: by applying the set of surface patterns offline, we were able to "precompile" from the AQUAINT corpus a list of nuggets about every entity mentioned within it. In essence, we have automatically constructed an immense relational database containing nuggets distilled from every article in the corpus. The task of answering definition questions then becomes a simple lookup for the relevant term. This approach is similar in spirit to the work reported by Fleischman et al. (2003) and Mann (2002), except that our system benefits from a greater variety of patterns and answers a broader range of questions.

Our surface patterns operated both at the word and part-of-speech level. Rudimentary chunking, such as marking the boundaries of noun phrases, was performed by grouping words based on their part-of-speech tags. In total, we applied eleven surface patterns over the entire corpus—these are detailed in Table 1, with examples in Table 2.

Typically, surface patterns identify nuggets on the order of a few words. In answering definition questions, however, we decided to return responses that include additional context—there is evidence that contextual information results in higher-quality answers (Lin et al., 2003). To accomplish this, all nuggets were expanded around their center point to encompass one hundred characters. We found that this technique enhances the readability of the responses, because many nuggets seem odd and out of place without context.

The results of applying our surface patterns to the entire AQUAINT corpus—the target, pattern type, nugget, and source sentence—are stored in a relational database. To answer a definition question, the target is used to query for all relevant nuggets in the database.

2.3 Dictionary Lookup

Another component of our system for answering definition questions utilizes an existing Web-based dictionary—dictionary definitions often supply knowledge that can be directly exploited. Previous factoid ques-

Copular	A **fractal** *is* a pattern that is irregular, but self-similar at all size scales
Become	**Althea Gibson** *became* the first black tennis player to win a Wimbledon singles title
Verb	**Francis Scott Key** *wrote* “The Star-Spangled Banner”
Appositive	The **Aga Khan**, Spiritual Leader of the Ismaili Muslims
Occupation	steel magnate **Andrew Carnegie**
Parenthesis	**Alice Rivlin** *(*director of the Office of Management and Budget*)*
Also-known-as	special proteins, *known as* **enzymes** // **amitriptyline**, *also known as* Elavil
Also-called	amino acid *called* **phenylalanine**
Or	**caldera**, *or* cauldron-like cavity on the summit
Like	prominent human rights leaders *like* **Desmond Tutu**
Relative clause	**Solar cells** *which* currently produce less than one percent of global power supplies

Table 2: Example nuggets for each pattern. (target term in bold, textual landmark in italics, and nugget underlined)

tion answering systems have already demonstrated the value of semistructured resources on the Web (Lin and Katz, 2003); we believe that some of these resources can be similarly employed to answer definition questions.

The setup of the TREC evaluations requires every answer to be paired with a supporting document; therefore, a system cannot simply return the dictionary definition of a term as its response. To address this issue, we developed answer projection techniques to “map” dictionary definitions back onto AQUAINT documents. Similar techniques have been employed for factoid questions, for example, in (Brill et al., 2001).

We have constructed a wrapper around the Merriam-Webster online dictionary. To answer a question using this technique, keywords from the target term’s dictionary definition and the target itself are used as the query to Lucene, a freely-available open-source IR engine. Our system retrieves the top one hundred documents returned by Lucene and tokenizes them into individual sentences, discarding candidate sentences that do not contain the target term. The remaining sentences are scored by their keyword overlap with the dictionary definition, weighted by the inverse document frequency of each keyword. All sentences with a non-zero score are retained and shortened to one hundred characters centered around the target term, if necessary.

The following are two examples of results from our dictionary lookup component:

> **What is the vagus nerve?**
> *Dictionary definition:* either of the 10th pair of cranial nerves that arise from the medulla and supply chiefly the viscera especially with autonomic sensory and motor fibers
> *Projected answer:* The vagus nerve is sometimes called the 10th cranial nerve. It runs from the brain . . .
>
> **What is feng shui?**
> *Dictionary definition:* a Chinese geomantic practice in which a structure or site is chosen or configured so as to harmonize with the spiritual forces that inhabit it
> *Projected answer:* In case you’ve missed the feng shui bandwagon, it is, according to Webster’s, “a Chinese geomantic practice . . .

This strategy was inspired by query expansion techniques often employed in document retrieval—essentially, the dictionary definition of a term is used as the source of expansion terms. Creative use of Web-based resources combined with proven information retrieval techniques enables this component to provide high quality responses to definition questions.

2.4 Document Lookup

If no answers are found by the previous two techniques, as a last resort our system employs traditional document retrieval to extract relevant nuggets. The target term is used as a Lucene query to gather a set of one hundred candidate documents. These documents are tokenized into individual sentences, and all sentences containing the target term are retained as responses (ranked by the Lucene-generated score of the document from which they came). These sentences are also shortened if necessary.

2.5 Answer Merging

The answer merging component of our system is responsible for integrating results from all three sources: database lookup, dictionary lookup, and document lookup. As previously mentioned, responses extracted using document lookup are used only if the other two methods returned no answers.

Redundancy presents a major challenge for integrating knowledge from multiple sources. This problem is especially severe for nuggets stored in our database. Since we precompiled knowledge about every entity instance in the entire AQUAINT corpus, common nuggets are often repeated. In order to deal with this problem, we applied a simple heuristic to remove duplicate information: if two responses share more than sixty percent of their keywords, one of them is randomly discarded.

After duplicate removal, all responses are ordered by the expected accuracy of the technique used to extract the nugget. To determine this expected accuracy, we performed a fine-grained evaluation for each surface pattern as well as the dictionary lookup strategy; we discuss these results further in Section 3.1.

Finally, the answer merging component decides how many responses to return. Given *n* total responses, we calculate the final number of responses to return as:

$$\begin{array}{ll} n & \text{if } n \leq 10 \\ n + \sqrt{n-10} & \text{if } n > 10 \end{array}$$

Having described the architecture of our system, we proceed to present evaluation results.

3 Evaluation

In this section we present two separate evaluations of our system. The first is a component analysis of our database and dictionary techniques, and the second involves our participation in the TREC 2003 Question Answering Track.

3.1 Component Evaluation

We evaluated the performance of each individual surface pattern and the dictionary lookup technique on 160 definition questions selected from the TREC-9 and TREC-10 QA Track testsets. Since we primarily generated our patterns by directly analyzing the corpus, these questions can be considered a blind testset. The performance of our surface patterns and our dictionary lookup technique is shown in Table 3.

Overall, database lookup retrieved approximately eight nuggets per question at an accuracy nearing 40%; dictionary lookup retrieved about 1.5 nuggets per question at an accuracy of 45%. Obviously, recall of our techniques is extremely hard to measure directly; instead, we use the prevalence of each pattern as a poor substitute. As shown in Table 3, some patterns occur frequently (e.g., e1_is and e1_appo), but others are relatively rare, such as the relative clause pattern, which yielded only six nuggets for the entire testset.

These results represent a baseline for the performance of each technique. Our focus was not on perfecting each individual pattern and the dictionary matching algorithm, but on building a complete working system. We will discuss future improvements and refinements in Section 5.

3.2 TREC 2003 Results

Our system for answering definition questions was independently and formally evaluated at the TREC 2003 Question Answering Track. For the first time, TREC evaluated definition questions in addition to factoid and list questions. Although our entry handled all three types of questions, we only report the results of the definition questions here; see (Katz et al., 2003) for description of the other components.

Pattern	**accuracy**	**nuggets**
e2_also	85.71	7
e2_aka	80.00	5
e2_occu	69.35	62
e1_or	67.74	31
e1_wdt	66.67	6
e2_like	64.60	113
e2_appo	60.00	20
e1_aka	50.00	2
e1_is	35.37	246
e1_pare	34.91	106
e1_appo	30.40	579
e1_verb	26.09	92
e1_beca	25.00	8
average	38.37	98.2
total		1277
dictionary	45.23	241

Table 3: Performance of each surface pattern and the dictionary lookup technique for all 160 test questions.

Group	**Run**	**F-measure**
MIT	MITCSAIL03a	0.309
	MITCSAIL03b	0.282
	MITCSAIL03c	0.282
Overall	best	0.555
	baseline IR	0.493
	median	0.192

Table 4: Official TREC 2003 results.

Overall, our system performed well, ranking eighth out of twenty-five groups that participated (Voorhees, 2003). Our official results for the definition sub-task are shown in Table 4, along with overall statistics for all groups. The formula used to calculate the F-measure is given in Figure 1. The β value of five indicates that recall is considered five times more important than precision, an arbitrary value set for the purposes of the evaluation.

Nugget precision is computed based on a length allowance of one hundred non-whitespace characters per relevant response, because a pilot study demonstrated that it was impossible for assessors to consistently enumerate the total set of “concepts” contained in a system response (Voorhees, 2003). The assessors’ nugget list (i.e., the ground truth) was created by considering the union of all responses returned by all participants. All relevant nuggets are divided into “vital” and “non-vital” categories, where vital nuggets are items of information that

Let r # of vital nuggets returned in a response
a # of non-vital nuggets returned in a response
R total # of vital nuggets in the assessors' list
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 $F(\beta = 5) = \dfrac{(\beta^2 + 1) \times \mathcal{P} \times \mathcal{R}}{\beta^2 \times \mathcal{P} + \mathcal{R}}$

Figure 1: Official definition of F-measure.

must be in a definition for it to be considered "good". Non-vital nuggets may also provide relevant information, but a "good" definition does not need to include them. Nugget recall is thus only a function of vital nuggets.

The best run, with an F-measure of 0.555, was submitted by BBN (Xu et al., 2003). The system used many of the same techniques we described here, with one important exception—they did not precompile nuggets into a database. In their own error analysis, they cited recall as a major cause of bad performance; this is an issue specifically addressed by our approach.

Interestingly, Xu et al. also reported an IR baseline which essentially retrieved the top 1000 sentences in the corpus that mentioned the target term (subjected to simple heuristics to remove redundant answers). This baseline technique achieved an F-measure of 0.493, which beat all other runs (expect for BBN's own runs). Because the F-measure heavily favored recall over precision, simple IR techniques worked extremely well. This issue is discussed in Section 4.1.

To identify areas for improvement, we analyzed the questions on which we did poorly and found that many of the errors can be traced back to problems with target extraction. If the target term is not correctly identified, then all subsequent modules have little chance of providing relevant nuggets. For eight questions, our system did not identify the correct target. The presence of stopwords and special characters in names was not anticipated:

What is Bausch & Lomb?
Who is Vlad the Impaler?
Who is Akbar the Great?

Our naive pattern-based parser extracted *Lomb*, *Impaler*, and *Great* as the target terms for the above questions. Fortunately, because *Lomb* and *Impaler* were rare terms, our system did manage to return relevant nuggets. However, since *Great* is a very common word, our nuggets for *Akbar the Great* were meaningless.

The system's inability to parse certain names is related to our simple assumption that the final consecutive sequence of capitalized words in a question is likely to be the target. This simply turned out to be an incorrect assumption, as seen in the following questions:

Who was Abraham in the Old Testament?
What is ETA in Spain?
What is Friends of the Earth?

Our parser extracted *Old Testament*, *Spain*, and *Earth* as the targets for these questions, which directly resulted in the system's failure to return relevant nuggets.

Our target extractor also had difficulty with apposition. Given the question "What is the medical condition shingles?", the extractor incorrectly identified the entire phrase *medical condition shingles* as the target term. Finally, our policy of ignoring articles before the target term caused problems with the question "What is the Hague?" Since we extracted *Hague* as the target term, we returned answers about a British politician as well as the city in Holland. Our experiences show that while target extraction seems relatively straightforward, there are instances where a deeper linguistic understanding is necessary.

Overall, our database and dictionary lookup techniques worked well. For six questions (out of fifty), however, neither technique found any nuggets, and therefore our system resorted to document lookup.

4 Evaluation Reconsidered

This section takes a closer look at the setup of the definition question evaluation at TREC 2003. In particular, we examine three issues: the scoring metric, error inherent in the evaluation process, and variations in judgments.

4.1 The Scoring Metric

As defined, nugget recall is only a function of the nuggets considered "vital". This, however, leads to a counter-intuitive situation where a system that returned every non-vital nugget but no vital nuggets would receive a score of zero. This certainly does not reflect the information needs of a real user—even in the absence of "vital" information, related knowledge might still be useful to a user. One solution might be to assign a relative weight to distinguish vital and non-vital nuggets.

The distinction between vital and non-vital nuggets is itself somewhat arbitrary. Consider some relevant nuggets for the question "What is Bausch & Lomb?":

world's largest eye care company
about 12000 employees
in 50 countries

Run	Total Nuggets	Relevant Returned	Recall
official	407	118	28.99%
fixed	407	120	29.48%

Table 5: Nugget recall, disregarding the distinction between vital and non-vital nuggets.

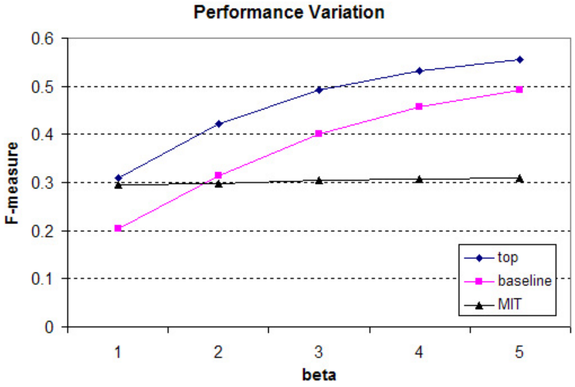

Figure 2: F-measure as a function of β.

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. We disagree and also believe that "based in Rochester, New York" is more important than "in 50 countries". Since it appears that the difference between vital and non-vital cannot be easily operationalized, there is little hope for systems to learn and exploit this distinction.

As a reference, we decided to reevaluate our system, ignoring the distinction between vital and non-vital nuggets. The overall nugget recall is reported in Table 5. We also report the nugget recall of our system after fixing our target extractor to handle the variety of target terms in the testset (the "fixed" run). Unfortunately, our performance for the fixed run did not significantly increase because the problem associated with unanticipated targets extended beyond the target extractor. Since our surface patterns did not handle these special entities, the database did not contain relevant entries for those targets.

Another important issue in the evaluation concerns the value of β, the relative importance between precision and recall in calculating the F-measure. The top entry achieved an F-measure of 0.555, but the response length averaged 2059 non-whitespace characters per question. In contrast, our run with an F-measure of 0.309 averaged only 620 non-whitespace characters per answer (only two other runs in the top ten had average response lengths lower than ours; the lowest was 338). Figure 2 shows F-measure of our system, the top run, and the IR baseline plotted against the value of β. As can be seen, if precision and recall are considered equally important (i.e., $\beta = 1$), the difference in performance between our system and that of the top system is virtually indistinguishable (and our system performs significantly better than the IR baseline). At the level of $\beta = 5$, it is obvious that standard IR technology works very well. The advantages of surface patterns, linguistic processing, answer fusion, and other techniques become more obvious if the F-measure is not as heavily biased towards recall.

What is the proper value of β? As this was the first formal evaluation of definition questions, the value was set arbitrarily. However, we believe that there is no "correct" value of β. Instead, the relative importance of precision and recall varies dramatically from application to application, depending on the user information need. A college student writing a term paper, for example, would most likely value recall highly, whereas the opposite would be true for a user asking questions on a PDA. We believe that these tradeoffs are worthy of further research.

4.2 Evaluation Error

In the TREC 2003 evaluation, we submitted three identical runs, but nevertheless received different scores for each of the runs. This situation can be viewed as a probe into the error margin of the evaluation—assessors are human and naturally make mistakes, and to ensure the quality of the evaluation we need to quantify this variation. Voorhees' analysis (2003) revealed that scores for pairs of identical runs differed by as much as 0.043 in F-measure.

For the three identical runs we submitted, there was one nugget missed in our first run that was found in the other two runs, ten nuggets from six questions missed for our second run that were found in the other runs, and ten nuggets from five questions missed in our third run. There were also nine nuggets from seven questions that were missed for all three runs, even though they were clearly present in our answers.

Together over our three runs, there were 48 nuggets from 13 questions that were clearly present in our responses but were not consistently recognized by the assessors. The question affected most by these discrepancies was "Who is Alger Hiss?", for which we received an F-measure of 0.671 in our first run, while for the second and third runs we received a score of zero.

If the 48 missed nuggets had been recognized by the assessors, our F-measure would be 0.327, 0.045 higher than the score we actually received for runs *b* and *c*. This single-point investigation is not meant to contest the relative rankings of submitted runs, but simply to demonstrate the magnitude of the human error currently present in the evaluation of definition questions (presumably, all groups suffered equally from these variations).

4.3 Variations in Judgment

The answers to definition questions were judged by humans, and humans naturally have differing opinions as to the quality of a response. These differences of opinion are not mistakes (unlike the issues discussed in the previous section), but legitimate variations in what assessors consider to be acceptable. These variations are compounded by the small size of the testset—only fifty questions. In a post-evaluation analysis, Voorhees (2003) determined that a score difference of at least 0.1 in F-measure is required in order for two evaluation results to be considered statistically different (at 95% confidence). A range of ±0.1 around our F-measure of 0.309 could either push our results up to fifth place or down to eleventh place.

A major source of variation is whether or not a passage matches a particular nugget in the assessor's list (the ground truth). Obviously, the assessors are not merely doing a string comparison, but are instead performing a "semantic match" of the relevant concepts involved. The following passages were rejected as matches to the assessors' nuggets:

Who is Al Sharpton?
Nugget: Harlem civil rights leader
Our answer: New York civil rights activist

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

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

Consider the nugget for Al Sharpton: although an "activist" may not be a "leader", and someone from New York may not necessarily be from Harlem, one might argue that the two nuggets are "close enough" to warrant a semantic match. The same situation is true of the other two questions. The important point here is that different assessors may judge these nuggets differently, contributing to detectable variations in score.

Another important issue is the composition of the assessors' nugget list, which serves as "ground truth". To insure proper assessment, each nugget should ideally represent an "atomic" concept—which in many cases, it does not. Again consider the nugget for Al Sharpton; "a Harlem civil rights leader" includes the concepts that he was an important civil rights figure and that he did much of his work in Harlem. It is entirely conceivable that a response would provide one fact but not the other. How then should this situation be scored? As another example, one of the nuggets for Alexander Pope is "English poet", which is clearly two separate facts.

Another desirable characteristic of the assessor's nugget list is uniqueness—nuggets should be unique, not only in their text but also in their meaning. In the TREC 2003 testset, three questions had exact duplicate nuggets. Furthermore, there were also several questions for which multiple nuggets are nearly synonymous (or are implied by other nuggets), such as the following:

What is TB?
highly infectious lung disease
contagious respiratory disease
common communicable disease

Who is Allen Iverson?
professional basketball player
philadelphia 76 er

What is El Shaddai?
catholic charismatic group
christian organization
catholic sect
religious group

Because the nuggets overlap greatly with each other in the concepts they denote, consistent and reproducible evaluation results are difficult.

Another desirable property of the ground truth is completeness, or coverage of the nuggets—which we also found to be lacking. There were many relevant items of information returned by our runs that did not make it onto the assessors' nugget list (even as non-vital nuggets). For the question "Who is Alberto Tomba?", the fact that he is Italian was not judged to be relevant. For "What are fractals?", the ground truth does not contain the idea that they can be described by simple formulas, which is one of their most important characteristics. Some more examples are shown below:

Aga Khan is the founder and principal shareholder of the Nation Media Group.
The vagus nerve is the sometimes known as the 10th cranial nerve.
Alexander Hamilton was an author, a general, and a founding father.
Andrew Carnegie established a library system in Canada.
Angela Davis taught at UC Berkeley.

This coverage issue also points to a deeper methodological problem with evaluating definition questions by pooling the results of all participants. Vital nuggets may be excluded simply because no system returned them. Unfortunately, there is no easy way to quantify this phenomenon.

Clearly, evaluating answers to definition questions is a challenging task. Nevertheless, consistent, repeatable, and meaningful scoring guidelines are critical to driving the development of the field. We believe that lessons learned from our analysis can lead to a more refined evaluation in the coming years.

5 Future Work

The results of our work highlight several areas for future improvement. As mentioned earlier, target extraction is a key, non-trivial capability critical to the success of a system. Similarly, database lookup works only if the relevant target terms are identified and indexed while preprocessing the corpus. Both of these issues point to the need for a more robust named-entity extractor, capable of handling specialized names (e.g., "Bausch & Lomb", "Destiny's Child", "Akbar the Great"). At the same time, the named-entity extractor must not be confused by sentences such as "*Raytheon & Boeing* are defense contractors" or "She gave *John the Honda* for Christmas".

Another area for improvement is the accuracy of the surface patterns. In general, our patterns only used local information; we expect that expanding the context on which these patterns operate will reduce the number of false matches. As an example, consider our e1_is pattern; in one test, over 60% of irrelevant nuggets were cases where the target is the object of a preposition and not the subject of the copular verb immediately following it. For example, this pattern matched the question "What is mold?" to the sentence "tools you need to look for mold are . . .". If we endow our patterns with better linguistic notions of constituency, we can dramatically improve their precision. Another direction we are pursuing is the use of machine learning techniques to learn predictors of good nuggets, much like the work of Fleischman et al. (2003). Separating "good" from "bad" nuggets fits very naturally into a binary classification task.

6 Conclusion

In this paper, we have described a novel set of strategies for answering definition questions from multiple sources: a database of nuggets precompiled offline using surface patterns, a Web-based electronic dictionary, and documents retrieved using traditional information retrieval technology. We have also demonstrated how answers derived using multiple strategies can be smoothly integrated to produce a final set of answers. In addition, our analyses have shown the difficulty of evaluating definition questions and inability of present metrics to accurately capture the information needs of real-world users. We believe that our research makes significant contributions toward the understanding of definition questions, a largely unexplored area of question answering.

7 Acknowledgement

This work was supported in part by the ARDA's Advanced Question Answering for Intelligence (AQUAINT) Program.

References

Eric Brill, Jimmy Lin, Michele Banko, Susan Dumais, and Andrew Ng. 2001. Data-intensive question answering. In *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*.

Michael Fleischman, Eduard Hovy, and Abdessamad Echihabi. 2003. Offline strategies for online question answering: Answering questions before they are asked. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*.

Ulf Hermjakob, Abdessamad Echihabi, and Daniel Marcu. 2002. Natural language based reformulation resource and Web exploitation for question answering. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*.

Boris Katz, Jimmy Lin, Daniel Loreto, Wesley Hildebrandt, Matthew Bilotti, Sue Felshin, Aaron Fernandes, Gregory Marton, and Federico Mora. 2003. Integrating Web-based and corpus-based techniques for question answering. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*.

Jimmy Lin and Boris Katz. 2003. Question answering from the Web using knowledge annotation and knowledge mining techniques. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM 2003)*.

Jimmy Lin, Dennis Quan, Vineet Sinha, Karun Bakshi, David Huynh, Boris Katz, and David R. Karger. 2003. What makes a good answer? The role of context in question answering. In *Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction (INTERACT 2003)*.

Gideon Mann. 2002. Fine-grained proper noun ontologies for question answering. In *Proceedings of the SemaNet'02 Workshop at COLING 2002 on Building and Using Semantic Networks*.

Martin M. Soubbotin and Sergei M. Soubbotin. 2001. Patterns of potential answer expressions as clues to the right answers. In *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*.

Ellen M. Voorhees. 2002. Overview of the TREC 2002 question answering track. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*.

Ellen M. Voorhees. 2003. Overview of the TREC 2003 question answering track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*.

Jinxi Xu, Ana Licuanan, and Ralph Weischedel. 2003. TREC2003 QA at BBN: Answering definitional questions. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*.