Answering What-Is Questions by Virtual Annotation

John Prager
IBM T.J. Watson Research Center
Yorktown Heights, N.Y. 10598
(914) 784-6809
jprager@us.ibm.com

Dragomir Radev
University of Michigan
Ann Arbor, MI 48109
(734) 615-5225
radev@umich.edu

Krzysztof Czuba
Carnegie-Mellon University
Pittsburgh, PA 15213
(412) 268 6521
kczuba@cs.cmu.edu

ABSTRACT
We present the technique of Virtual Annotation as a specialization of Predictive Annotation for answering definitional **What is** questions. These questions generally have the property that the type of the answer is not given away by the question, which poses problems for a system which has to select answer strings from suggested passages. Virtual Annotation uses a combination of knowledge-based techniques using an ontology, and statistical techniques using a large corpus to achieve high precision.

Keywords
Question-Answering, Information Retrieval, Ontologies

1. INTRODUCTION
Question Answering is gaining increased attention in both the commercial and academic arenas. While algorithms for general question answering have already been proposed, we find that such algorithms fail to capture certain subtleties of particular types of questions. We propose an approach in which different types of questions are processed using different algorithms. We introduce a technique named Virtual Annotation (VA) for answering one such type of question, namely the **What is** question.

We have previously presented the technique of Predictive Annotation (PA) [Prager, 2000], which has proven to be an effective approach to the problem of Question Answering. The essence of PA is to index the semantic types of all entities in the corpus, identify the desired answer type from the question, search for passages that contain entities with the desired answer type as well as the other query terms, and to extract the answer term or phrase. One of the weaknesses of PA, though, has been in dealing with questions for which the system cannot determine the correct answer type required. We introduce here an extension to PA which we call Virtual Annotation and show it to be effective for those "What is/are (a/an) X" questions that are seeking hypernyms of X. These are a type of definition question, which other QA systems attempt to answer by searching in the document collection for textual clues similar to those proposed by [Hearst, 1998], that are characteristic of definitions. Such an approach does not use the strengths of PA and is not successful in the cases in which a deeper understanding of the text is needed in order to identify the defining term in question.

We first give a brief description of PA. We look at a certain class of **What is** questions and describe our basic algorithm. Using this algorithm we develop the Virtual Annotation technique, and evaluate its performance with respect to both the standard TREC and our own benchmark. We demonstrate on two question sets that the precision improves from .15 and .33 to .78 and .83 with the addition of VA.

2. BACKGROUND
For our purposes, a question-answering (QA) system is one which takes a well-formed user question and returns an appropriate answer phrase found in a body of text. This generally excludes **How** and **Why** questions from consideration, except in the relatively rare cases when they can be answered by simple phrases, such as "by fermenting grapes" or "because of the scattering of light". In general, the response of a QA system will be a named entity such as a person, place, time, numerical measure or a noun phrase, optionally within the context of a sentence or short paragraph.

The core of most QA systems participating in TREC [TREC8, 2000 & TREC9, 2001] is the identification of the answer type desired by analyzing the question. For example, **Who** questions seek people or organizations, **Where** questions seek places, **When** questions seek times, and so on. The goal, then, is to find an entity of the right type in the text corpus in a context that justifies it as the answer to the question. To achieve this goal, we have been using the technique of PA to annotate the text corpus with semantic categories (QA-Tokens) prior to indexing.

Each QA-Token is identified by a set of terms, patterns, or finite-state machines defining matching text sequences. Thus "Shakespeare" is annotated with "PERSON$", and the text string "PERSON$" is indexed at the same text location as "Shakespeare". Similarly, "$123.45" is annotated with "MONEY$". When a question is processed, the desired QA-Token is identified and it replaces the Wh-words and their auxiliaries. Thus, "Who" is replaced by "PERSON$", and "How much" + "cost" are replaced by "MONEY$". The resulting query is then input to the search engine as a bag of words. The expectation here is that if the initial question were "Who wrote Hamlet", for example, then the modified query of "PERSON$ write Hamlet" (after lemmatization) would be a

perfect match to text that states "Shakespeare wrote Hamlet" or "Hamlet was written by Shakespeare".

The modified query is matched by the search engine against passages of 1-2 sentences, rather than documents. The top 10 passages returned are processed by our Answer Selection module which re-annotates the text, identifies all potential answer phrases, ranks them using a learned evaluation function and selects the top 5 answers (see [Radev et al., 2000]).

The problem with "What is/are (a/an) X" questions is that the question usually does not betray the desired answer type. All the system can deduce is that it must find a noun phrase (the QA-Token THING$). The trouble with THING$ is that it is too general and labels a large percentage of the nouns in the corpus, and so does not help much in narrowing down the possibilities. A second problem is that for many such questions the desired answer type is not one of the approximately 50 high-level classes (i.e. QA-Tokens) that we can anticipate at indexing; this phenomenon is seen in TREC9, whose 24 definitional **What is** questions are listed in the Appendix. These all appear to be calling out for a hypernym. To handle such questions we developed the technique of *Virtual Annotation* which is like PA and shares much of the same machinery, but does not rely on the appropriate class being known at indexing time. We will illustrate with examples from the animal kingdom, including a few from TREC9.

3. VIRTUAL ANNOTATION

If we look up a word in a thesaurus such as WordNet [Miller et al., 1993]), we can discover its hypernym tree, but there is no indication which hypernym is the most appropriate to answer a **What is** question. For example, the hypernym hierarchy for "nematode" is shown in Table 1. The level numbering counts levels up from the starting term. The numbers in parentheses will be explained later.

Table 1. Parentage of "nematode" according to WordNet.

Level	Synset
0	{nematode, roundworm}
1	{worm(13)}
2	{invertebrate}
3	{animal(2), animate being, beast, brute, creature, fauna}
4	{life form(2), organism(3), being, living thing}
5	{entity, something}

At first sight, the desirability of the hypernyms seems to decrease with increasing level number. However, if we examine "meerkat" we find the hierarchy in Table 2.

We are leaving much unsaid here about the context of the question and what is known of the questioner, but it is not unreasonable to assert that the "best" answer to "What is a meerkat" is either "a mammal" (level 4) or "an animal" (level 7). How do we get an automatic system to pick the right candidate?

Table 2. Parentage of "meerkat" according to WordNet

Level	Synset
0	{meerkat, mierkat}
1	{viverrine, viverrine mammal}
2	{carnivore}
3	{placental, placental mammal, eutherian, eutherian mammal}
4	{mammal}
5	{vertebrate, craniate}
6	{chordate}
7	{animal(2), animate being, beast, brute, creature, fauna}
8	{life form, organism, being, living thing}
9	{entity, something}

It seems very much that what we would choose intuitively as the best answer corresponds to Rosch et al.'s *basic categories* [Rosch et al., 1976]. According to psychological testing, these are categorization levels of intermediate specificity that people tend to use in unconstrained settings. If that is indeed true, then we can use online text as a source of evidence for this tendency. For example, we might find sentences such as "… meerkats and other Y …", where Y is one of its hypernyms, indicating that Y is in some sense the preferred descriptor.

We count the co-occurrences of the target search term (e.g. "meerkat" or "nematode") with each of its hypernyms (e.g. "animal") in 2-sentence passages, in the TREC9 corpus. These counts are the parenthetical numbers in Tables 1 and 2. The absence of a numerical label there indicates zero co-occurrences. Intuitively, the larger the count, the better the corresponding term is as a descriptor.

3.1 Hypernym Scoring and Selection

Since our ultimate goal is to find passages describing the target term, discovering zero co-occurrences allows elimination of useless candidates. Of those remaining, we are drawn to those with the highest counts, but we would like to bias our system away from the higher levels. Calling a nematode a life-form is correct, but hardly helpful.

The top levels of WordNet (or any ontology) are by definition very general, and therefore are unlikely to be of much use for purposes of definition. However, if none of the immediate parents of a term we are looking up co-occur in our text corpus, we clearly will be forced to use a more general term that does. We want to go further, though, in those cases where the immediate parents do occur, but in small numbers, and the very general parents occur with such high frequencies that our algorithm would select them. In those cases we introduce a tentative level ceiling to prevent higher-level terms from being chosen if there are suitable lower-level alternatives.

We would like to use a weighting function that decreases monotonically with level distance. Mihalcea and Moldovan [1999], in an analogous context, use the logarithm of the number of terms in a given term's subtree to calculate weights, and they claim to have shown that this function is optimal. Since it is approximately true that the level population increases

exponentially in an ontology, this suggests that a linear function of level number will perform just as well.

Our first step is to generate a *level-adapted count* (LAC) by dividing the co-occurrence counts by the level number (we are only interested in levels 1 and greater). We then select the best hypernym(s) by using a fuzzy maximum calculation. We locate the one or more hypernyms with greatest LAC, and then also select any others with a LAC within a predefined threshold of it; in our experimentation we have found that a threshold value of 20% works well. Thus if, for example, a term has one hypernym at level 1 with a count of 30, and another at level 2 with a count of 50, and all other entries have much smaller counts, then since the LAC 25 is within 20% of the LAC 30, both of these hypernyms will be proposed.

To prevent the highest levels from being selected if there is any alternative, we tentatively exclude them from consideration according to the following scheme:

If the top of the tree is at level N, where N <= 3, we set a tentative ceiling at N-1, otherwise if N<=5, we set the ceiling at N-2, otherwise we set the ceiling at N-3. If no co-occurrences are found at or below this ceiling, then it is raised until a positive value is found, and the corresponding term is selected.

If no hypernym at all co-occurs with the target term, then this approach is abandoned: the "What" in the question is replaced by "THING$" and normal procedures of Predictive Annotation are followed.

When successful, the algorithm described above discovers one or more candidate hypernyms that are known to co-occur with the target term. There is a question, though, of what to do when the question term has more than one sense, and hence more than one ancestral line in WordNet. We face a choice of either selecting the hypernym(s) with the highest overall score as calculated by the algorithm described above, or collecting together the best hypernyms in each parental branch. After some experimentation we made the latter choice. One of the questions that benefitted from this was "What is sake". WordNet has three senses for sake: good (in the sense of welfare), wine (the Japanese drink) and aim/end, with computed scores of 122, 29 and 87/99 respectively. It seems likely (from the phrasing of the question) that the "wine" sense is the desired one, but this would be missed entirely if only the top-scoring hypernyms were chosen.

We now describe how we arrange for our Predictive Annotation system to find these answers. We do this by using these descriptors as *virtual* QA-Tokens; they are not part of the search engine index, but are tagged in the passages that the search engine returns at run time.

3.2 Integration

Let us use H to represent either the single hypernym or a disjunction of the several hypernyms found through the WordNet analysis. The original question Q =

"What is (a/an) X"

is converted to Q' =

"DEFINE$ X H"

where DEFINE$ is a *virtual* QA-Token that was never seen at indexing time, does not annotate any text and does not occur in the index. The processed query Q' then will find passages that contain occurrences of both X and H; the token DEFINE$ will be ignored by the search engine. The top passages returned by the search engine are then passed to Answer Selection, which re-annotates the text. However, this time the virtual QA-Token DEFINE$ is introduced and the patterns it matches are defined to be the disjuncts in H. In this way, all occurrences of the proposed hypernyms of X in the search engine passages are found, and are scored and ranked in the regular fashion. The end result is that the top passages contain the target term and one of its most frequently co-occurring hypernyms in close proximity, and these hypernyms are selected as answers.

When we use this technique of Virtual Annotation on the aforementioned questions, we get answer passages such as

> *"Such genes have been found in nematode worms but not yet in higher animals."*

and

> *"South African golfer Butch Kruger had a good round going in the central Orange Free State trials, until a mongoose-like animal grabbed his ball with its mouth and dropped down its hole. Kruger wrote on his card: "Meerkat.""*

4 RESULTS

4.1 Evaluation

We evaluated Virtual Annotation on two sets of questions – the definitional questions from TREC9 and similar kinds of questions from the Excite query log (see http://www.excite.com). In both cases we were looking for definitional text in the TREC corpus. The TREC questions had been previously verified (by NIST) to have answers there; the Excite questions had no such guarantee. We started with 174 Excite questions of the form "What is X", where X was a 1- or 2-word phrase. We removed those questions that we felt would not have been acceptable as TREC9 questions. These were questions where:

- The query terms did not appear in the TREC corpus, and some may not even have been real words (e.g. "What is a gigapop").[1] 37 questions.
- The query terms were in the corpus, but there was no definition present (e.g "What is a computer monitor").[2] 18 questions.
- The question was not asking about the class of the term but how to distinguish it from other members of the class (e.g. "What is a star fruit"). 17 questions.
- The question was about computer technology that emerged after the articles in the TREC corpus were written (e.g. "What is a pci slot"). 19 questions.
- The question was very likely seeking an example, not a definition (e.g. "What is a powerful adhesive"). 1 question plus maybe some others – see the Discussion

[1] That is, after automatic spelling correction was attempted.

[2] The TREC10 evaluation in August 2001 is expected to contain questions for which there is no answer in the corpus (deliberately). While it is important for a system to be able to make this distinction, we kept within the TREC9 framework for this evaluation.

section later. How to automatically distinguish these cases is a matter for further research.

Of the remaining 82 Excite questions, 13 did not have entries in WordNet. We did not disqualify those questions.

For both the TREC and Excite question sets we report two evaluation measures. In the TREC QA track, 5 answers are submitted per question, and the score for the question is the reciprocal of the rank of the first correct answer in these 5 candidates, or 0 if the correct answer is not present at all. A submission's overall score is the mean reciprocal rank (MRR) over all questions. We calculate MRR as well as mean binary score (MBS) over the top 5 candidates; the binary score for a question is 1 if a correct answer was present in the top 5 candidates, 0 otherwise. The first sets of MBS and MRR figures are for our base system, the second set the system with VA.

Table 3. Comparison of base system and system with VA on both TREC9 and Excite definitional questions.

Source	No. of Questions	MBS w/o VA	MRR w/o VA	MBS with VA	MRR with VA
TREC9 (in WN)	20	.3	.2	.9	.9
TREC9 (not in WN)	4	.5	.375	.5	.5
TREC9 Overall	**24**	**.333**	**.229**	**.833**	**.833**
Excite (in WN)	69	.101	.085	.855	.824
Excite (not in WN)	13	.384	.295	.384	.295
Excite Overall	**82**	**.146**	**.118**	**.780**	**.740**

We see that for the 24 TREC9 definitional questions, our MRR score with VA was the same as the MBS score. This was because for each of the 20 questions where the system found a correct answer, it was in the top position.

By comparison, our base system achieved an overall MRR score of .315 across the 693 questions of TREC9. Thus we see that with VA, the average score of definitional questions improves from below our TREC average to considerably higher. While the percentage of definitional questions in TREC9 was quite small, we shall explain in a later section how we plan to extend our techniques to other question types.

4.2 Errors

The VA process is not flawless, for a variety of reasons. One is that the hierarchy in WordNet does not always exactly correspond to the way people classify the world. For example, in WordNet a dog is not a pet, so "pet" will never even be a candidate answer to "What is a dog".

When the question term is in WordNet, VA succeeds most of the time. One of the error sources is due to the lack of uniformity of the semantic distance between levels. For example, the parents of "architect" are "creator" and "human", the latter being what our system answers to "What is an architect". This is technically correct, but not very useful.

Another error source is polysemy. This does not seem to cause problems with VA very often – indeed the co-occurrence calculations that we perform are similar to those done by [Mihalcea and Moldovan, 1999] to perform word sense disambiguation – but it can give rise to amusing results. For example, when asked "What is an ass" the system responded with "Congress". *Ass* has four senses, the last of which in WordNet is a slang term for sex. The parent synset contains the archaic synonym *congress* (uncapitalized!). In the TREC corpus there are several passages containing the words *ass* and *Congress*, which lead to *congress* being the hypernym with the greatest score. Clearly this particular problem can be avoided by using orthography to indicate word-sense, but the general problem remains.

5 DISCUSSION AND FURTHER WORK

5.1 Discussion

While we chose not to use Hearst's approach of key-phrase identification as the primary mechanism for answering **What is** questions, we don't reject the utility of the approach. Indeed, a combination of VA as described here with a key-phrase analysis to further filter candidate answer passages might well reduce the incidence of errors such as the one with *ass* mentioned in the previous section. Such an investigation remains to be done.

We have seen that VA gives very high performance scores at answering **What is** questions – and we suggest it can be extended to other types – but we have not fully addressed the issue of automatically selecting the questions to which to apply it. We have used the heuristic of only looking at questions of the form "What is (a/an) X" where X is a phrase of one or two words. By inspection of the Excite questions, almost all of those that pass this test are looking for definitions, but some - such as "What is a powerful adhesive" - very probably do not. There are also a few questions that are inherently ambiguous (understanding that the questioners are not all perfect grammarians): is "What is an antacid" asking for a definition or a brand name? Even if it is known or assumed that a definition is required, there remains the ambiguity of the state of knowledge of the questioner. If the person has no clue what the term means, then a parent class, which is what VA finds, is the right answer. If the person knows the class but needs to know how to distinguish the object from others in the class, for example "What is a star fruit", then a very different approach is required. If the question seems very specific, but uses common words entirely, such as the Excite question "What is a yellow spotted lizard", then the only reasonable interpretation seems to be a request for a subclass of the head noun that has the given property. Finally, questions such as "What is a nanometer" and "What is rubella" are looking for a value or more common synonym.

5.2 Other Question Types

The preceding discussion has centered upon **What is** questions and the use of WordNet, but the same principles can be applied to other question types and other ontologies. Consider the question "Where is Chicago", from the training set NIST supplied for TREC8. Let us assume we can use statistical arguments to decide that, in a vanilla context, the question is about the city as opposed to the rock group, any of the city's sports teams or the University. There is still considerable ambiguity regarding the granularity of the desired answer. Is it: Cook County? Illinois? The Mid-West? The United States? North America? The Western Hemisphere? …

There are a number of geographical databases available, which either alone or with some data massaging can be viewed as ontologies with "located within" as the primary relationship. Then by applying Virtual Annotation to **Where** questions we can find the enclosing region that is most commonly referred to in the context of the question term. By manually applying our algorithm to "Chicago" and the list of geographic regions in the previous paragraph we find that "Illinois" wins, as expected, just beating out "The United States". However, it should be mentioned that a more extensive investigation might find a different weighting scheme more appropriate for geographic hierarchies.

The aforementioned answer of "Illinois" to the question "Where is Chicago?" might be the best answer for an American user, but for anyone else, an answer providing the country might be preferred. How can we expect Virtual Annotation to take this into account? The "hidden variable" in the operation of VA is the corpus. It is assumed that the user belongs to the intended readership of the articles in the corpus, and to the extent that this is true, the results of VA will be useful to the user.

Virtual Annotation can also be used to answer questions that are seeking examples or instances of a class. We can use WordNet again, but this time look to hyponyms. These questions are more varied in syntax than the **What is** kind; they include, for example from TREC9 again:

"Name a flying mammal."
"What flower did Vincent Van Gogh paint?"

and

"What type of bridge is the Golden Gate Bridge?"

6. SUMMARY

We presented Virtual Annotation, a technique to extend the capabilities of PA to a class of definition questions in which the answer type is not easily identifiable. Moreover, VA can find text snippets that do not contain the regular textual clues for presence of definitions. We have shown that VA can considerably improve the performance of answering **What is** questions, and we indicate how other kinds of questions can be tackled by similar techniques.

7. REFERENCES

[1] Hearst, M.A. "Automated Discovery of WordNet Relations" in *WordNet: an Electronic Lexical Database*, Christiane Fellbaum Ed, MIT Press, Cambridge MA, 1998.

[2] Mihalcea, R. and Moldovan, D. "A Method for Word Sense Disambiguation of Unrestricted Text". *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pp. 152-158, College Park, MD, 1999.

[3] Miller, G. "WordNet: A Lexical Database for English", *Communications of the ACM* 38(11) pp. 39-41, 1995.

[4] Moldovan, D.I. and Mihalcea, R. "Using WordNet and Lexical Operators to Improve Internet Searches", *IEEE Internet Computing*, pp. 34-43, Jan-Feb 2000.

[5] Prager, J.M., Radev, D.R., Brown, E.W. and Coden, A.R. "The Use of Predictive Annotation for Question-Answering in TREC8", *Proceedings of TREC8*, Gaithersburg, MD, 2000.

[6] Prager, J.M., Brown, E.W., Coden, A.R., and Radev, D.R. "Question-Answering by Predictive Annotation", *Proceedings of SIGIR 2000*, pp. 184-191, Athens, Greece, 2000.

[7] Radev, D.R., Prager, J.M. and Samn, V. "Ranking Suspected Answers to Natural Language Questions using Predictive Annotation", *Proceedings of ANLP'00*, Seattle, WA, 2000.

[8] Rosch, E. et al. "Basic Objects in Natural Categories", *Cognitive Psychology* 8, pp. 382-439, 1976.

[9] TREC8 - "The Eighth Text Retrieval Conference", E.M. Voorhees and D.K. Harman Eds., NIST, Gaithersburg, MD, 2000.

[10] TREC9 - "The Ninth Text Retrieval Conference", E.M. Voorhees and D.K. Harman Eds., NIST, Gaithersburg, MD, to appear.

APPENDIX

What-is questions from TREC9

617: What are chloroplasts? (X)
528: What are geckos?
544: What are pomegranates?
241: What is a caldera? (X)
358: What is a meerkat?
434: What is a nanometer? (X)
354: What is a nematode?
463: What is a stratocaster?
447: What is anise?
386: What is anorexia nervosa?
635: What is cribbage?
300: What is leukemia?
305: What is molybdenum?
644: What is ouzo?
420: What is pandoro? (X)
228: What is platinum?
374: What is porphyria?
483: What is sake?
395: What is saltpeter?
421: What is thalassemia?
438: What is titanium?
600: What is typhoid fever?
468: What is tyvek?
539: What is witch hazel?

Our system did not correctly answer the questions marked with an "X". For all of the others the correct answer was the first of the 5 attempts returned.