Use of WordNet Hypernyms for Answering What-Is Questions

John Prager, Jennifer Chu-Carroll
IBM T.J. Watson Research Center
Yorktown Heights, N.Y. 10598
jprager/jencc@us.ibm.com

Krzysztof Czuba[1]
Carnegie-Mellon University
Pittsburgh, PA 15213
kczuba@cs.cmu.edu

Abstract

We present a preliminary analysis of the use of WordNet hypernyms for answering "What-is" questions. We analyse the approximately 130 definitional questions in the TREC10 corpus with respect to our technique of Virtual Annotation (VA), which has previously been shown to be effective on the TREC9 definitional question set and other questions. We discover that VA is effective on a subset of the TREC10 definitional questions, but that some of these questions seem to need a user model to generate correct answers, or at least answers that agree with the NIST judges. Furthermore, there remains a large enough subset of definitional questions that cannot benefit at all from the WordNet isa-hierarchy, prompting the need to investigate alternative external resources.

1. Introduction

Work in the field of Question-Answering has taken off since the introduction of a QA track in TREC in 1999 (see, e.g. [Voorhees and Tice, 2000]). Much of the published work in the field has centered around the fact-based questions that form the current basis of this track. While differing greatly in the specifics, most of the systems published in the literature to date use a similar approach (at the coarsest level of description) of a sequence of processing stages: the question is analysed to discover the answer-type that is sought, a query is constructed from the question (with appropriate vocabulary expansions and morphological normalization), a standard IR search is performed, documents or passages are retrieved and these texts are examined for presence of terms of the appropriate answer type, possibly in a context that satisfies other derived criteria (see for example [Clarke et al. 2001, Ittycheriah et al. 2001, Moldovan et al. 2000, Prager et al. 2000, Srihari and Li 2000]). Some systems, such as Falcon [Harabagiu et al., 2001] and some of our own experimental prototypes, are using feedback loops to retry stages that are deemed unsuccessful.

One recurring question type is the definitional question, usually of the form "What is/are <noun phrase>", although other syntactic forms are used but with essentially the same meaning. The difficulty that arises with these questions is that the answer type is left completely open. Even the Webclopedia system [Hovy et al., 2001], which employs an extensive question typology, cannot be very specific with these questions. The TREC9 question set consisted of about 5% definitional questions, while the TREC10 set, which appears to better mirror actual user questions (Ellen Voorhees, personal communication), consisted of about 26% definitional. Thus we believe that examining what is required to answer this kind of question is worthwhile.

Granted, there are many occasions where the text explicitly provides a definition with sentences of the form "X is <something>" – in fact by a cursory analysis of the judgment sets some 82% of the TREC10 definitional questions are answered by copular expressions. However, relying on this is easily seen to be problematic. Firstly, definitions are provided using overall a wide variety of syntactic structures, but more importantly, very sophisticated NLP is required to determine that the <something> above is a definition rather than some arbitrary predicate. Clearly some additional component is required. WordNet [Miller, 1995] is currently the preferred resource for ontological information, and promises to be very helpful for this particular problem. We have previously shown [Prager et al. 2001] its effectiveness for a small class of "What-is" questions; in this paper we examine the effectiveness of the WordNet

[1] Work performed while at the IBM T.J. Watson Research Center.

hypernym (or "isa") hierarchy for the TREC10 "What-is" questions.

2. Predictive/Virtual Annotation for Question-Answering

Our Question-Answering system employs the technique of Predictive Annotation, introduced and described in [Prager et al. 2000a]. The technique revolves around the concept of semantic class labels that we call QA-Tokens, corresponding loosely to some of the Basic Categories of [Rosch et al. 1976]. These are used not only as Named Entity descriptors, but are actual tokens processed by the indexer. For example, people are tagged with PERSON$, lengths of time with DURATION$. For named entity detection we use Textract [Wacholder, Ravin and Choi, 1997, Byrd and Ravin, 1999], and for search we use Guru-QA, based on Guru [Brown and Chong, 1997], but with a specialized weighting scheme and ranking algorithm.

Identifying the semantic answer-type (QA-Token set) in the question (e.g. "Who" -> PERSON$ and "How long" -> DURATION$ or LENGTH$) and matching against a semantically tagged corpus only works if such information is conveyed by the question either explicitly or implicitly. Questions beginning with "Who", "When" and "Where" fulfill this requirement, as do those with "How + <adjective>" or "How + <adverb>", and also "What (or Which) + <noun phrase>". However, definitional "What-is" questions (e.g., "What is a nematode?") do not indicate the answer type, thus rendering the annotations in the corpus ineffective.

For such questions we need an alternative approach. One possibility is to find all occurrences of the question term in the corpus, and to analyze all these documents (or at least the passages surrounding the instances) for key terms or phrases indicating a definition, as did Hearst [1998], and Joho and Sanderson [2000]. However, we have adopted another approach, more in line with our disposition to shift the computational burden in the direction of IR rather than NLP. As described in [Prager et al. 2001], this approach has been shown to give an accuracy of 83% on TREC9 "What-is" questions. This sample set was rather small (24 questions) and was thus not a reliable indicator of its general efficacy.

Our approach stems from the observations that (1) providing the parent class should be a good answer to a definitional "What-is" question, and (2) frequently terms are encountered in text along with their class (e.g. "nematodes and other worms", "metals such as tungsten", "gecko (a lizard)", and so on). WordNet is a good, easily-accessible ontological resource for finding the isa-hierarchy of a term, and so we use WordNet to find the best class descriptor(s) for the question term and include them as additional search terms.

Our WordNet lookup algorithm works by counting co-occurrences of the question term with each of its WordNet ancestors in the TREC corpus, and dividing this number by the number of isa-links between the two. The best terms, by this calculation, win. This approach guarantees that the selected terms co-occur[2] with the question term, and therefore that answer passages can be found.[3]

Since our search process ([Prager et al. 2000]) is passage-based, we look for short passages that contain both the question term and any of its ancestors that our WordNet lookup algorithm proposes. According to criteria such as described in [Radev at al 2000, Chu-Carroll et al. in progress], the best answer fragments are returned.

3. Performance Evaluation and Data Analysis

While our algorithm was shown to be very effective on TREC9 "What-is" questions, it was much less so on TREC10. Hence we decided to examine the assumptions inherent in the process in order to understand more fully the conditions under which our algorithm is effective.

The assumptions underlying our approach were as follows:

1. The question term is in WordNet.
2. At least one of its ancestors is useful as a definition.
3. Such ancestors (in #2) are themselves sufficient as definitions.
4. Our algorithm can find the ancestor(s).

We need to explain the distinction between conditions #2 and #3. We have found that there are some cases where an ancestor provides a definition that would best be extended by further qualification on the ancestral term, e.g., by citing the difference between the term and others in its ancestral class.[4] For example, saying that

[2] within a two-sentence passage.

[3] In a small number of cases, the question term is present in WordNet but none of the ancestors co-occur with it anywhere in the TREC corpus.

[4] We realize that it is a subjective decision as to whether or not a term makes for an acceptable definition. We have made

an amphibian (TREC10 #944) is an animal is technically correct, but it is considerably more useful to say that it is an animal that lives both on land and in water. (It can also be a vehicle, but the same analysis holds.) Thus, just calling an amphibian an animal violates assumption #3. However, we maintain that, even though "animal" by itself does not provide for a sufficiently useful answer for "amphibian", including the search term "animal" will likely lead us to passages in which good definitions for amphibian can be found. On the other hand, we have found that there are terms for which none of the ancestors are particularly useful, even as partial definitions. For example, the parentage of "eclipse" (TREC10 #1016) in WordNet is the synset-chain: {interruption, break, abrupt change}, {happening, occurrence, natural event}, {event}, while a good definition would talk about one astronomical body blocking or obscuring another. In other words, one cannot easily make a simple definition by adding premodifiers or prepositional phrases to the ancestral noun. For questions like this one, assumption #2 is violated.

For the purpose of analyzing the effectiveness of our algorithm, we identified 130 TREC10 questions which sought the definition of a given term or phrase. Although most of these questions are phrased in the "What is/are X?" format, we included those questions that were similar in nature, such as "What does X mean?" and "What does X do?" Since WordNet includes a small number of famous people, we also process "Who is/was X" questions in the same way.

Granting that there is occasional subjectivity involved, we have grouped the 130 definitional questions into 5 groups according to which of assumptions 1-4 have been violated. More specifically, questions are classified based on the following criteria:

?? Group 1: question term is not in WordNet.

?? Group 2: no hypernym is particularly useful as part of a definition.

?? Group 3: "best" ancestor is useful as a partial definition, but needs to be further qualified.

?? Group 4: "best" ancestor is sufficient as a definition for the question term by itself. But our WordNet lookup algorithm failed to return it as the best candidate.

?? Group 5: "best" ancestor is a good definition by itself and our algorithm found it.

Table 1 shows a summary of relevant statistics for the 5 groups,[5] while Table 2 - Table 6 contain detailed information about each group used to generate the summary. MRR is Mean Reciprocal Rank of the first correct answer and is in the range 0-1.

Group	Count	MRR
1	25	0.171
2	19	0.097
3	40	0.283
4	14	0.232
5	32	0.812

Table 1 Summary of Question Classification

Table 2 - Table 6 consist of the following columns: the TREC10 question number, the question term, what our algorithm finds as a suitable ancestor (possibly a disjunction), and the score our system receives (given as rank of first correct answer). Note that this score "r" is based not on our run as submitted to NIST, but after fixing a bug that was later found; where the fixed system differed from the original, evaluation was done by reference to the NIST-supplied judgment sets. Question terms in italics are those for which NIST asserts there was no answer in the TREC corpus.

Trec#	Question Term	WordNet Ancestor(s)	r
915	biosphere		0
947	fibromyalgia		0
961	spider veins		0
997	Duke Ellington		1
1022	Wimbledon		5
1026	target heart rate		0
1034	severance pay		0
1042	Phi Beta Kappa		0
1051	nanotechnology		5
1075	neuropathy		5
1077	cryptography		0
1114	ozone depletion		0
1116	Sitting Shiva		0
1141	*home equity*		0
1148	pilates		0
1160	dianetics		0
1180	pulmonary fibrosis		0
1185	foot and mouth disease		0
1262	Moulin Rouge		2
1267	mad cow disease		0

every attempt to base our analysis on the TREC10 judgment set whenever possible.

[5] The MRR for each group is calculated by disregarding those questions known to have NIL as answers.

1289	die-casting		0
1324	bangers and mash		0
1330	spirometer test		1
1385	bio-diversity		0
1393	e-coli		1

Table 2 Group 1: Question Term Not in WordNet

Table 2 enumerates those 25 questions in which the question term is not present in WordNet. For these questions, our system does not benefit from the virtual annotation mechanism, and, as a result, found answers to only 7 of them within the 50-byte answer fragments using our default mechanism.

Trec#	Question Term	WordNet Ancestor(s)	r
897	atom	{particle, matter, molecule}	0
903	autism	{syndrome}	2
974	prism	{form, optical prism}	0
985	desktop publishing	{business}	0
992	coral reefs	{formation}	0
1016	eclipse	{break}	0
1033	platelets	{blood platelet}	1
1046	sunspots	{point}	0
1054	*obtuse angle*	*	0
1088	relative humidity	*	0
1121	spleen	{ire, anger, tissue}	0
1135	Valentine's Day	*	0
1170	viscosity	{property}	0
1179	antacids	{cause}	4
1243	acid rain	{acid precipitation}	0
1255	ciao	{message}	0
1273	annuity	{payment}	0
1303	metabolism	{activity}	0
1363	compounded interest	{cost, charge}	0

Table 3 Group 2: No Hypernym Forms Useful Part of Definition

In Table 3, we show 19 questions for which none of the ancestors of the question term in WordNet are particularly useful even as partial definitions. The third column in the table shows what our WordNet lookup algorithm proposes as the "best" ancestor,[6] although to classify a question into this category, we have manually examined the other ancestors to ensure that our algorithm did not overlook other suitable candidates.

Table 2 and Table 3 contain a total of 44 questions, or about 1/3 of all definitional questions, for which WordNet's utility in aiding question answering is minimal at best. This fact is further confirmed by the statistics shown in Table 1, where the MRR scores for groups 1 and 2 are substantially lower than those for the other groups. This prompts the need to investigate other supplemental sources of information for when the WordNet isa-hierarchy fails. In addition, although it is obvious when additional information is needed for those questions in Table 2, it is not a trivial task for a system to determine when an ancestor proposed by WordNet is unlikely to be found in the definition of a question term and should therefore be discarded. We leave the investigation of both of these issues as future work.

Trec#	Question Term	WordNet Ancestor(s)	r
918	cholesterol	{alcohol}	0
920	caffeine	{compound}	0
926	invertebrates	{animal}	0
935	Teflon	{plastic}	2
944	amphibian	{vehicle, amphibious vehicle, animal, aircraft}	0
969	pH scale	{measure}	1
982	xerophytes	{plant, planting}	0
991	cryogenics	{science, field}	0
994	neurology	{study, medicine}	0
1005	acupuncture	{treatment}	1
1028	*foreclosure*	{proceeding, proceed}	0
1043	nicotine	{substance}	4
1055	polymers	{compound}	0
1067	supernova	{star}	1
1102	defibrillator	{device}	0
1108	fungus	{plant, planting}	0
1129	sonar	{device}	2
1131	phosphorus	{element}	1
1138	bandwidth	{measure}	0
1140	parasite	{organism, leech, sponge}	1
1142	meteorologist	{expert, specialist}	0
1152	Mardi Gras	{carnival, day}	3
1166	osteoporosis	{health problem}	1
1169	esophagus	{passage}	0
1192	barometer	{instrument}	0
1196	solar wind	{radiation}	0
1209	fuel cell	{device}	1
1214	diabetes	{disorder}	4

[6] In those questions marked by an asterisk, our algorithm did not return any candidate term because none of the question term's WordNet ancestors co-occur with it in the TREC corpus.

1258	acetic acid	{compound}	5
1266	pathogens	{microorganism}	1
1285	carcinogen	{substance}	3
1288	nepotism	{favoritism}	3
1300	carbon dioxide	{compound, CO}	3
1309	semiconductors	{semiconductor device, material}	0
1310	nuclear power	{energy}	0
1322	enzymes	{protein}	0
1362	solar cells	{photovoltaic cell}	0
1365	antigen	{drug}	0
1370	thermometer	{instrument}	0
1384	pectin	{sugar}	0

Table 4 Group 3: "Best" Hypernym Not Specific Enough as Definition

Table 4 shows 40 questions where WordNet proposes an ancestor which requires further qualification (either in the form of a premodifier or a prepositional phrase postmodifier) in order to constitute a useful definition. For example, “cholesterol” can be defined as a “fatty alcohol” and “invertebrates” as “animals without backbones.” Column three in the table again shows the ancestor returned by our WordNet lookup algorithm, which is included as part of at least one NIST-judged correct answer in each case.

Note that the Virtual Annotation algorithm we originally described in [Prager et al. 2001] looked strictly at ancestor terms in the isa-hierarchy. Synonyms were only examined when explicitly called for by questions of the form “What is another name for X”. However, following the observation that sometimes in “What is X” questions the “X” is a rare synonym for a better-known term, in this experiment we treated the question-term’s synset as a level-0 parent. This backfired when it initially found “oesophagus” as the meaning of “esophagus”, for example, and “grippe” for “influenza”, but we found that in general it was more helpful to include the synset of the question term in the analysis. Testing for orthographic or other such variations helped eliminate the former kind of problem, and filtering on occurrence count ratios the latter.

Trec#	Question Term	WordNet Ancestor(s) found/could have been found	r
912	epilepsy	{disorder}/ {neurological disorder}	1
917	bipolar disorder	{condition}/ {manic depression}	0
1081	leukemia	{cancer}/ {cancer of the blood}	1
1113	influenza	{disease}/ {contagious disease}	4
1159	fortnight	{period}/ {two weeks}	0
1183	strep throat	{disease}/ {sore throat}	0
1188	Aborigines	{}/ {(original) inhabitant}	0
1207	pneumonia	{disease}/ {respiratory disease}	0
1224	mold	{plant}/ {fungus}	2
1248	quicksilver	{substance, matter}/ {mercury}	0
1280	Muscular Dystrophy	{disease}/ {genetic disorder, genetic disease}	0
1317	genocide	{kill, killing}/ {racial extermination}	0
1377	rheumatoid arthritis	{disease}/ {inflammatory disease}	0
1379	cerebral palsy	{disorder}/ {nervous disorder}	2

Table 5 Group 4: "Best" Ancestor Makes Good Definition But Was Not Found

Table 5 illustrates 14 examples in which there exists a better WordNet ancestor than the one proposed by our lookup algorithm. The third column in the table shows two sets of terms, the first of which is the term selected by our algorithm and the second of which is a term also present in the WordNet hierarchy that we prefer over the selected term as a definition of the question term. In all cases, the selected term is a hypernym of the preferred term, which has a very low or zero co-occurrence count with the question term. In addition, note that in many cases, the selected term consists of the head noun of the preferred term, which includes an additional adjectival premodifier or a prepositional phrase postmodifier.

As discussed earlier, our question answering system includes the proposed WordNet hypernym as an additional search term for passage retrieval. For questions in groups 3 and 4, this means that the search is biased toward passages that include terms that could potentially form a definition for the question term. The effect of the inclusion of such terms is evidenced by the statistics in Table 1, where the MRRs for groups 3 and 4 are higher than for groups 1 and 2, which received no help for WordNet at all. However, the improvement in MRR scores is less than we would have liked. We plan to investigate more sophisticated answer-selection mechanisms for identifying contexts in which definitions are provided.

Trec#	Question Term	WordNet Ancestor(s)	r
896	Galileo	{astronomer}	1
936	amitriptyline	{antidepressant}	1
937	shaman	{priest}	1
959	Abraham Lincoln	{(frontier) lawyer}	1
980	amoxicillin	{antibiotic}	1
999	micron	{micrometer}	0
1038	poliomyelitis	{infantile paralysis}	1
1044	vitamin B1	{thiamine}	1
1058	Northern Lights	{aurora borealis}	0
1061	acetaminophen	{painkiller}	1
1110	sodium chloride	{salt}	1
1126	phenylalanine	{amino acid}	1
1137	hypertension	{high blood pressure}	1
1168	peyote	{mescaline mescalin mescal}	2
1177	chunnel	{Chunnel Tunnel}	1
1181	Qaaludes	{methaqualone}	2
1182	naproxen	{drug, anti-inflammatory}	2
1223	Milky Way	{galaxy}	1
1230	semolina	{flour}	1
1232	Ursa Major	{constellation}	1
1254	thyroid	*{thyroid gland}*	0
1271	ethics	{study, morality}	2
1282	propylene glycol	{antifreeze}	1
1283	panic disorder	{anxiety disorder}	1
1290	myopia	{nearsightedness}	1
1311	tsunami	{wave, tidal wave}	1
1320	earthquake	*{temblor}*	0
1328	ulcer	{ulceration}	1
1329	vertigo	{dizziness}	1
1352	schizophrenia	{mental illness}	1
1360	pediatricians	{baby doctor}	1
1364	capers	{pickle}	1

Table 6 Group 5: Best Found Hypernym Makes Good Definition

The final group of definitional questions, shown in Table 6, contains those where the ancestor proposed by our algorithm constitutes a useful definition of the question term by itself.[7] Not surprisingly, for this group of questions, our system returned the correct definition in the first position in the vast majority of cases, and as a result received a very high MRR score, as shown in Table 1.

4. Discussion

The issue of what constitutes a correct answer has raged in the TREC community since the first QA track in TREC8, and shows no sign of being settled. One particularly important but neglected issue is that of knowing who the questioner is. In everyday communication, people ask questions of each other, and in all cases the answers given are conditioned on the responder's knowledge of the questioner and suspicions of what they know, what they don't know and how much they are seeking to learn.

For argument's sake, one can postulate several different kinds of questioner. These might include: a child, an intelligent adult for whom English (or in the case of TREC10 #1255 "What does ciao mean", Italian) is or is not their primary language, or a student learning a new field (so he might well know other technical terms in the field). NIST has not asserted any user model. Unfortunately, it is not that easy to induce one from the judgment sets made available. It is particularly difficult to infer what level of specificity is required in an answer. For example, carbon dioxide is not a compound (according to NIST) yet nanotechnology is a science; diabetes is not a disorder yet acupuncture is a treatment, influenza is not a disease but poliomyelitis is.

Given a user description, it should be straightforward to determine the correct answer level; at least it should give rise to less haphazard specificity levels of correct answer. For instance, consider TREC10 #1266: "What are xerophytes". For all but botanists or landscape gardeners, the answer "plants" is probably sufficient, absent any context.

One approach that might be worth taking is to generate alternative answers based on the different user-model assumptions, and to assume a priori probabilities of these different models. These probabilities can be fixed, or (outside of TREC) determined by exterior processes. Within the TREC paradigm, however, one can possibly infer something about the questioner from the question itself. An average intelligent adult could very reasonably ask "What are xerophytes?", but maybe not so reasonably ask "What is the Milky Way?". Even the article in the question can convey meaning: "What is a thyroid?" might well be asked by a child, but "What is the thyroid?" might be asked by an anatomy professor of a medical student (i.e. the definite article here can convey tacit agreement of the domain, in this case the

[7] Here the effect of subject judgment comes into play. Those WordNet ancestors in italics were not considered correct answers by the NIST judges.

human body, which might be all that is needed to answer a child).

The difficulty with paying close attention to the question syntax is that if indeed the question is asked by a child or a non-native speaker, then conclusions based on correct grammaticality may be unreliable. "What is mold?" (TREC10 #1224) requires a very different answer from "What is a mold?", but only if presence or absence of the indefinite article can be trusted; if we knew the question came with a Russian accent, for example, we would have more information to work with! Answering the question properly requires identifying an appropriate user model. Doing this requires, in part, analysis of the question syntax. Drawing valid conclusions from the question syntax again requires a user model!

5. Conclusions and Future Work

We have broken down the 130 TREC10 definitional questions into five groups according to how useful an algorithm that seeks primarily to define a term by its WordNet class or genus can be. We have ideas about how to address each group, but the challenge is in identifying which situation is present for any particular question.

Our system had the worst performance with groups 1 and 2, when a term was not in WordNet or it had an entry but its WordNet parents were not useful for definitional purposes – in fact the latter case fared worse than the former because our system was distracted into thinking it had an answer. One possible solution is to manually explicitly identify these general hypernyms (property, cause, activity etc.) and to make our program try another approach if these are initially proposed.

The next-ranking groups (nos. 3 and 4) were those where the located WordNet ancestor was promising but not specific enough, and where our algorithm selected a non-optimal ancestor. The former problem can possibly be addressed by selecting (to add to the search) significant terms from the WordNet gloss in the hopes that they are differentiae of the genus. The latter problem can in individual cases be fixed by retuning the parameters in our lookup algorithm, but we don't want the successful cases (primarily in group 5) to start failing. It is unclear right now for how large a subset of groups 4 and 5 a successful parameter set can be found.

Group 5 fared very well, which gives us hope that WordNet will be useful in the future for a significant number of definitional questions – our groups 3-5 totalled two-thirds of the TREC10 set.

We have not had time to explore those cases in group 5 where we did not find the right answer, according to the NIST assessors, nor why the definitions in our group 3 were considered not specific enough. For many of these cases, arguments can be made that, depending on who asked the question, the right answer was found. A more complete analysis requires both a model of the user and of what constitutes a good answer to a question. We hope to pursue this line of inquiry in the near future.

References

[1] Brown, E.W. and Chong, H.A. "The Guru System in TREC-6." *Proceedings of TREC6*, Gaithersburg, MD, 1998.

[2] Byrd, R. and Ravin, Y. "Identifying and Extracting Relations in Text", *Proceedings of NLDB 99*, Klagenfurt, Austria, 1999.

[3] Chu-Carroll, J., Prager, J., Ravin, Y., Cesar, C. "A Hybrid Approach to Natural Language Web Queries", in preparation.

[4] Clarke, C.L.A., Cormack, G.V., Kisman, D.I.E., Lynam, T.R. Question Answering by Passage Selection. In *Proceedings of the 9th Text Retrieval Conference (TREC9)*, Gaithersburg, MD, to appear in 2001.

[5] Harabagiu, S., Moldovan, D., Pasca, M., Mihalcea, R., Surdeanu, M., Bunescu, R., Gîrju, R., Rus, V., Morarescu, P. FALCON: Boosting Knowledge for Answer Engines. In *Proceedings of the 9th Text Retrieval Conference (TREC9)*, Gaithersburg, MD, to appear in 2001.

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

[7] Hovy, H., Gerber, L., Hermjakob, U., Lin, Chin-Yew, Ravichandran, D. "Towards Semantic-Based Answer Pinpointing", *Proceedings of Human Language Technologies Conference*, pp. 339-345, San Diego CA, March 2001

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

[9] Ittycheriah, A., Franz, M., Zhu, W-J., Ratnaparkhi, A., Mammone, R.J. IBM's Statistical Question Answering System, In *Proceedings of the 9th Text Retrieval Conference (TREC9)*, Gaithersburg, MD, to appear in 2001.

[10] Joho, H and Sanderson, M. “Retrieving Descriptive Phrases from Large amounts of Free Text”, *Proceedings of CIKM, 2000*.

[11] Moldovan, D., Harabagiu, S., Pasca, M., Mihalcea, R., Girju, R., Goodrum, R. and Rus, V. The Structure and Performance of an Open-Domain Question Answering System. In *Proceedings of the Conference of the Association for Computational Linguistics* (ACL-2000), 563–570.

[12] 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.

[13] 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.

[14] Prager, J.M., Radev, D.R. and Czuba, K. “Answering What-Is Questions by Virtual Annotation.” *Proceedings of Human Language Technologies Conference*, San Diego CA, pp. 26-30, March 2001

[15] 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.

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

[17] Srihari, R. and W. Li. 2000. A Question Answering System Supported by Information Extraction. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics* (NAACL-00), 166–172.

[18] Voorhees, E.M. and Tice, D.M. “Building a Question Answering Test Collection”, *Proceedings of SIGIR 2000*, pp. 184-191, Athens, Greece, 2000.

[19] Wacholder, N., Ravin, Y. and Choi, M. “Disambiguation of Proper Names in Text”, *Proceedings of ANLP’97*. Washington, DC, April 1997.