Exploiting Paraphrases and Deferred Sense Commitment to Interpret Questions more Reliably

Peter Clark and Phil Harrison
Boeing Research & Technology
{peter.e.clark,philip.harrison}@boeing.com

Abstract

Creating correct, semantic representations of questions is essential for applications that can use formal reasoning to answer them. However, even within a restricted domain, it is hard to anticipate all the possible ways that a question might be phrased, and engineer reliable processing modules to produce a correct semantic interpretation for the reasoner. In our work on posing questions to a biology knowledge base, we address this brittleness in two ways: First, we exploit the DIRT paraphrase database to introduce alternative phrasings of a question; Second, we defer word sense and semantic role commitment until question answering. Resulting ambiguities are then resolved by interleaving additional interpretation with question-answering, allowing the combinatorics of alternatives to be controlled and domain knowledge to guide paraphrase and sense selection. Our evaluation suggests that the resulting system is able to understand exam-style questions more reliably.

1 Introduction

Our goal is to allow users to pose exam-style questions to a biology knowledge base (KB), containing formal representations of biological structures and processes expressed in first-order logic. As the questions typically require automated reasoning to answer them, a semantic interpretation of each question is needed. In our earlier work (Clark et al, 2007), questions were interpreted using a conventional pipeline (parse, coreference, sense and role disambiguation). However, despite moderate performance, the original ("base") system suffered from well-known problems of brittleness, arising from both premature commitments in the pipeline and the system's limited knowledge of the multiple ways that questions can be expressed. In this paper, we describe how deferred commitment and a large paraphrase database can be used to reduce these problems, drawing on prior work and applying it in the context of a large KB being available. In particular, by interleaving interpretation and answering, we are able to control the combinatorics of alternatives that would otherwise arise. An evaluation suggests that this improves the ability of the system to correctly interpret, and hence answer, questions.

2 Context and Related Work

Our system aims to interpret and answer high-school level, exam-style biology questions, expressed in sentence form. Our source of answers is a formal knowledge-base and reasoning engine (rather than a text corpus), placing specific requirements on the interpretation process - in particular, a full semantic interpretation of the question is required. Questions are typically one or two sentences long, for example:

(1) Does a prokaryotic cell contain ribosomes?
(2) A eukaryotic cell has a nucleus. Does that nucleus contain RRNA?
(3) Is adenine found in RNA molecules?
(4) Does a prokaryotic cell have a region consisting of cytosol?
(5) Do ribosomes synthesize proteins in the cytoplasm?
(6) What is the material, containing DNA and protein, that forms into chromosomes during mitosis?

Coling 2008: Poster Volume, pages 171–179,
Beijing, August 2010

Interpreting and answering this style of question has a long history in NLP, both for answers found via database retrieval and formal reasoning, and for answers extracted from a large text corpus.

For answers found using reasoning, the focus of this paper, early NL systems typically used a pipelined architecture for question interpretation (e.g., Bobrow, 1964; Woods 1977), with later systems also using semantic constraints to guide disambiguation decisions (e.g., Novak, 1977). More recently, as well as there being significant improvements in the performance of typical pipeline modules, e.g., word sense disambiguation (Navigli, 2009), there has been substantial work on various forms of deferred commitment, underspecification, and paraphrasing to expand the space of interpretations considered, and thus improve interpretation. Underspecified representations (e.g., van Deemter and Peters, 1996; Pinkal, 1999) allow ambiguity (in particular scope ambiguity) to be preserved in a single structure and commitments deferred until later, allowing multiple interpretations to be carried through the system. Similarly, a system can defer commitment by simply carrying multiple, alternative interpretations forward as individual structures, or packed together into a single structure (e.g., Alshawi and van Eijck, 1989, Bobrow et al., 2005; Kim et al., 2010a,b). Finally, canonicalized representations are often used to represent (and hence carry through the system) multiple, equivalent surface forms as a single structure, e.g., normalizing active and passive forms, or alternative forms of noun modification (Rinaldi et al., 2003). All these techniques help avoid premature commitment in interpretation.

As well as avoiding early rejection of interpretations in these ways, there has been substantial, recent work on expanding the space of possible interpretations considered through the use of paraphases (e.g., Sekine and Inui, 2007). Paraphrasing is based on the observation that there are many ways of saying (roughly) the same thing, and that syntatic manipulation alone is not sufficient to enumerate them all. Paraphrases aim to enumerate these additional alternatives, and may be generated synthetically (e.g., Rinaldi et al., 2003), drawn from similar texts (e.g., from similar questions for QA, Harabagiu et al., 2000), or mined from a corpus using machine learning techniques (e.g., Lin and Pantel, 2001). They have proved to be particularly useful in the context of textual entailment (e.g., Bentivogli et al., 2009), and in corpus-based question answering (e.g., Harabagiu et al., 2003).

Our work builds on this prior work, applying and extending these ideas to the context where a formal knowledge base and reasoning engine is available. In particular, we interleave the process of expanding the space of interpretations considered (using paraphrases and deferred commitment) with the process of question answering (which narrows down that space by selecting interpretations supported by the KB), thus controlling the otherwise combinatorial explosion of alternatives. This makes it feasible to use the DIRT paraphrase database (12 million paraphrases) for generating a full semantic interpretation of the original question, extending its previous use in the semi-formal context of textual entailment (Bentivogli et al., 2009). Our use of reasoning to guide disambiguation follows Hobbs et als (1993) method of "interpretation as abduction", where the system searches a space of possible interpretations for one(s) that are provable from the KB, preferring those interpretations.

3 The Problem

Although the biology KB we are using contains the knowledge to answer the six earlier questions (1)-(6), only the first two are correctly answered with the original pipelined ("base") system. For question (3):

(3) Is adenine found in RNA molecules?

the system (mis-)interprets this as referring to some actual "finding" event, not recognizing that this is an alternative way of phrasing a question about physical structure. Similarly, the notion of "consisting of" in question (4) is an unexpected phrasing that the system does not understand. Questions (5) and (6) are also answered incorrectly by the base system due to errors in semantic role labeling during interpretation. In (5):

(5) Do ribosomes synthesize proteins in the cytoplasm?

"in" is (mis-)interpreted by the language interpreter as an is-inside(x,y) relation, while the KB itself represents this relationship as site(x,y), hence the system fails to produce the correct answer (yes). Similarly, for (6) "into" is (mis)interpreted as destination(x,y) but represented in the KB as result(x,y).

Clearly, one can tweak the original interpreter to overcome these particular problems. However, it is a slow, expensive process, and in general it is impossible to anticipate all such problems up front. Statistical methods (e.g., Manning and Schutze, 1999) offer an alternative approach but one that is similarly noisy, problematic for question-answering applications.

4 Solution Approach

The brittleness of the base system can be partially attributed to its eager commitments, ahead of specifics that might be discovered during question-answering itself. To address this, we have modified the system in two ways. First, we have added use of paraphrases to explore additional interpretations of the question during question-answering. Second, we defer sense and semantic role disambiguation until question answering. As a result, part of interpretation occurs during answering itself: multiple interpretations are tried and a commitment is made to the one(s) that produce a non-null answer. The justification for this commitment is a **benevolent user** assumption, namely that the interpretation that "makes sense" with respect to the KB (i.e., produces a non-null answer) is the one that the user intended.

This use of question-answering to drive disambiguation follows Hobbs et als. (1993) work on Interpretation as Abduction. In that framework, a system searches for an interpretation that is provable from the KB plus a minimal cost set of assumptions, the interpretation corresponding to a particular way to disambiguate the text. In our work we do a similar thing, although restrict the assumptions to disambiguation decisions and exclude assuming new knowledge, as we are dealing with questions rather than assertions (if no interpretations are provable, then we treat the answer as "no" rather than treating the unproven query as something that should be asserted as true).

4.1 Paraphrases

Several paraphrase databases are now available to the NLP community[1], typically built by automatically finding phrases that occur in distributionally similar contexts (e.g., Dras et al, 2005). To date, paraphrase databases have primarily been exploited for recognizing textual entailment (e.g., Bentivogli et al., 2009, Clark et al, 2009), and for corpus-based question answering (e.g., Harabagiu et al., 2003). Here we use them for generating a full semantic interpretation in the context of querying a formal knowledge resource.

We use the DIRT paraphrase database (Lin and Pantel, 2001), containing approximately 12 million automatically learned rules of the form:

IF X *relation* Y THEN X *relation'* Y

where *relation* is a path in the dependency tree between constitutents X and Y, or equivalently (as we use later) a *chain of literals*:

$$\{p_0(x_0,x_1), w_1(x_1), \ldots p_{n-1}(x_{n-1},x_n)\}$$

where p_i is the syntactic relation between (non-prepositional) constituents x_i and x_{i+1}, and w_i is the word used for x_i. An example from DIRT is:

IF X is found in Y THEN X is inside Y

The condition "X is found in Y" can be expressed as the chain of literals:

{ object-of(x,f), "find"(f), "in"(f,y) }

The database itself is noisy, containing both good and nonsensical paraphrases. Interestingly, their use in question-answering tends to filter out most bad paraphrases, as it is rare that a nonsensical paraphrases will by chance produce an answer (i.e., the question + KB together help "triangulate" on good paraphrases). Nevertheless, bad paraphrases can sometimes produce incorrect answers. To handle this in a practical setting, we are adding an interactive interface (outside the scope of this paper) that shows the user any paraphrases used, and allows him/her to verify/block them as desired.

4.2 Deferred Sense Commitment

A second, common cause of failure of the base system was incorrect assignment of senses and

[1] e.g., http://www.aclweb.org/aclwiki/index.php?title=RTE_Knowledge_Resources

semantic relations during word sense disambiguation (WSD) and semantic role labeling (SRL). While domain-specific terms are generally reliably disambiguated, disambiguation of general terms (e.g., whether "split" denotes the concept of Separate or Divide) and semantic roles (e.g., whether "into" denotes destination(x,y) or result(x,y)) is less reliable, with only limited improvement attainable through manual engineering or machine learning. The problem is compounded by a degree of subjectivity in the way knowledge is encoded in the KB, for example whether the KB engineer chose to conceptualize a biological object as the "agent" or "instrument" or "site" of an activity is to a degree a matter of viewpoint.

To overcome this, we defer WSD and SRL commitments until question-answering itself. One can view this as a trivial form of preserving underspecification (eg. Pinkal, 1999) in the initial language processing, where the words themselves denote their possible meanings.

4.3 Algorithm and Implementation

Questions are first parsed using a broad coverage, phrase structure parser, followed by coreference resolution, producing an initial "syntactic" logical form, for example:

Question: *Do mitotic spindles consist of hollow microtubules?*

Logical Form (LF): "mitotic-spindle"(s), "consist"(c), "hollow"(h), "microtubule"(m), subject(c,s), "of"(c,m), modifier(m,h).

Next, rather than attempting word sense disambiguation (WSD) and semantic role labeling (SRL) as would be done in the base system, the system immediately starts work on answering the question, even though a complete semantic interpretation has not yet been produced. In the process of answering, the system explores alternative word senses, semantic roles, and paraphrases for the particular literals it is working on (described shortly), and if any are provable from the knowledge in the knowledge base then those branch(es) of the search are explored further. There are two basic steps in this process:

(a) **setup:** create an instance X0 of the object being universally quantified over[2] (identified during initial language interpretation)

(b) **query:** for each literal in the LF with at least one bound variable, iteratively query the KB to see if some interpretation of those literals are provable i.e., already known.

In this example, illustrated in Figure 1, for step (a) the system first creates an instance X0 of a mitotic spindle, i.e., asserts the instantiated first literal isa(X0,Mitotic-Spindle), and then queries the inference engine with the remaining LF literals. (If there are multiple senses for "mitotic spindle", then an instance for each sense is created, to be explored in parallel). For step (b), the system uses the algorithm as follows:

repeat
  select a chain C_u of "syntactic" literals in the LF with at least 1 bound variable
    C_u = {p(x,y)} or {w(x)} or
      {p_1(x,z), w(z), p_2(z,y)}
  select some interpretation C of C_u where:
    C is a possible interpretation of C_u
    or C'_u is a possible paraphrase for C_u and
      C is a possible interpretation of C'_u
  try prove C[bindings] → new-bindings
  If success:
    replace C_u with C
    add new-bindings to bindings
until
  all clauses are proved

where:

- A *syntactic literal* is a literal whose predicate is a word or syntactic role (subject, object, modifier, etc.) All literals in the initial LF are syntactic literals.
- A *chain of literals* is a set of syntactic literals in the LF of the form {p(x,y)} or {w(x)} or {p_1(x,z), w(z), p_2(z,y)}, where p_i, w are words or syntactic roles (subject, mod, etc).
- A *possible paraphrase* is a possible substitution of one chain of literals with another, listed in the DIRT paraphrase database.

[2] If the system can prove the answer for a (new) instance X0 of the universally quantified class, then it holds for all instances, i.e., if KB ∪ f(X0) ├ g(X0) then KB ├ f(X0)→g(X0), hence KB ├ ∀x f(x)→g(x) via the principle of universal generalization (UG).

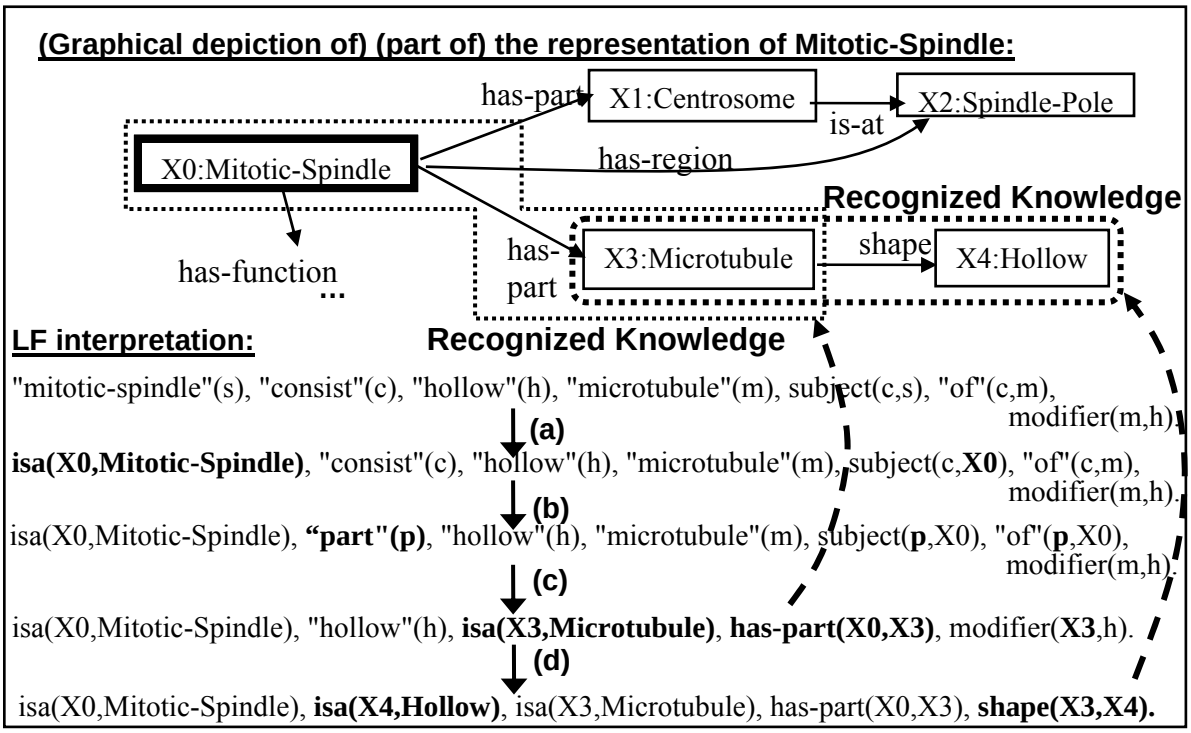

Figure 1: The path found through the search space for an interpretation of the example question. (a) setup (b) paraphrase substitution (IF X consists of Y THEN Y is part of X) (c) interpretation of {subject-of(X0,p),"part"(p), "of"(p,X0)} as has-part(X0,m), preferred as it is provable from the KB, resulting in m=X3 (d) interpretation of the syntactic modifier(X3,h) relation (from "hollow microtubule") as shape(X3,h) as it is provable from the KB.

- A *possible interpretation* of the singleton chain of literals {w(x)} is isa(x,*class*), where *class* is a possible sense of word w.
- A *possible interpretation* of a chain of literals {p(x,y)} or {p1(x,z),w(z),p2(z,y)} is r(x,y), where r is a semantic relation corresponding to syntactic relation p (e.g., "in"(x,y) → is-inside(x,y)) or word w (e.g., {subject-of(e,h), "have"(h), "of"(h,n)} → has-part(e,n)).

Possible word-to-class and word-to-predicate mappings are specified in the KB.

Figure 1 illustrates this procedure for the example sentence. The procedure iteratively replaces syntactic literals with semantic literals that correspond to an interpretation that is provable from the KB. If all the literals are proved, then the answer is "yes", as there exists an interpretation under which it can be proved from the KB, under the benevolent user assumption that this is the interpretation that the user intended.

As there are several points of non-determinism in the algorithm, e.g., which literals to select, which interpretation to explore, it is a search process. Our current implementation uses most-instantiated-first query ordering plus breadth-first search, although other implementations could traverse the space in other ways.

5 Evaluation

To evaluate the system, we measured its question-answering performance on a set of 141 true/false biology questions, ablating paraphrases and deferred commitment to measure their impact. The 141 questions were sentencized versions of the multiple choice options in 22 original AP-level exam questions that, in an earlier evaluation (Clark, 2009), users had difficulty rephrasing into a form that the system understood. Each original multiple choice option was minimally rewritten as a complete sentence (most multiple choice questions were partial se-

Configuration	Accuracy (score = y/y+n/n)	system/actual answers			
		y/y	n/y	y/n	n/n
Naive(all false)	67% (94)	0	47	0	94
Base system	72% (102)	8	41	0	94
+ Paraphrases	75% (106)	13	34	1	93
+ Deferred commitment	76% (107)	13	34	0	94
+ Both (full system)	84% (118)	25	22	1	93

Table 1: Performance of different configurations of the system. The y/y column shows the number of questions for which the system answered "yes" and the correct answer is "yes", etc.

ntences), while preserving the original English phrasing. For example the original question:

73. Which of the following best describes the DNA molecule?
   a. Two parallel strands of nitrogen bases held together by hydrogen bonding
   b. Two complementary strands of deoxyribose and phosphates held together by hydrogen bonding
   c. Two antiparallel strands of nucleotides held together by hydrogen bonding
   d. A single strand of nitrogen bases coiled upon itself by hydrogen bonding
   e. A single strand of nucleotides coiled into a helix.

was rewritten as five questions:

- Does a DNA molecule have two parallel strands of nitrogen bases held together by hydrogen bonding?
- Does a DNA molecule have two complementary strands of deoxyribose and phosphates held together by hydrogen bonding?
- Does a DNA molecule have two antiparallel strands of nucleotides held together by hydrogen bonding?
- Does a DNA molecule have a single strand of nitrogen bases coiled upon itself by hydrogen bonding?
- Does a DNA molecule have a single strand of nucleotides coiled into a helix?

Similarly:

79. All of the following organelles are associated with protein synthesis EXCEPT:
   a. ribosomes; b. Golgi bodies;...; e...

was rewritten as five questions:

- Are ribosomes associated with protein synthesis?
- Are Golgi bodies associated with...*etc*.

For 18 of the original questions, each of the 5 options expanded to 1 true/false question. For 3 comparison questions ("Which X is in Y but not Z?"), each option expanded into 2 questions ("Is X in Y?" "Is X in Z?"). Finally 1 question involved parallelism ("Which of the following A,B,C do X,Y,Z respectively?") which expanded into 21 questions ("Does A do X?" "Does A do Y?" etc.) after removing duplicates. Of the resulting 141 questions, 47 had the "gold" answer of true, 94 false. Of the 47 positives, 4 were out of scope of the reasoning engine, involving questions about possibility rather than truth, for example:

- Can a DNA adenine bond to an RNA uracil?

Another 3 were out of scope of the knowledge in the KB (2 requiring unrepresented temporal knowledge and 1 requiring commonsense knowledge). Thus the upper bound on performance, given the particular KB and reasoning engine that we are using, is 134/141 (95%).

We ran the base system alone, with paraphrasing (only), with deferred commitment (only), and with both. The results are shown in Table 1. As can be seen, true negatives (n/y) are a substantially larger challenge than false positives (y/n), as the system answers "no" by defalt if it is unable to prove the facts in the interpreted question from the KB. During interpretation, the base "pipeline" system commits to disambiguation decisions at each step, and if any commitment is wrong then it will also get the answer wrong, as reflected

by the only small (8) increase in number correctly answered.

Paraphrases allow the system to search for alternative interpretations, adding five more questions to be answered correctly but also introducing one false positive (y/n). The false positive was for the question:

Do peroxisomes make proteins?

This was (incorrectly) answered "yes" by the system as it used a bad DIRT paraphrase (IF X makes Y THEN X is made from Y), selected because it led to a provable interpretation (peroxisomes are made (synthesized) from proteins), but not the one the author intended. It is an interesting and perhaps somewhat surprising result that this was the only false positive, given that the DIRT database is noisy (approximately half its paraphrases are questionable or invalid). The low number of false positives appears to be due to the fact that the vast number of invalid paraphrases produce nonsensical, hence unprovable and rejected, interpretations.

Similarly, deferred commitment (alone) allowed five additional questions (different to those for paraphrasing) to be answered, again as premature word sense and semantic role labeling was avoided. For example, for "...the polymerase builds a strand...", the pipeline prematurely commits to the strand being the object of the build, while in the KB it is represented as the result of the build. Deferred commitment allows the system to search and find such alternatives.

Finally there were several (7) questions requiring both paraphrases and deferred commitment to answer. For example, "Do mitochondria provide cellular energy?" was answered using both a paraphrase (IF X provides Y THEN X creates Y) and deferred commitment (mitochondria was correctly interpreted as the site of the creation, as represented in the KB, while the pipeline prematurely committed to agent).

Although deferring SRL and WSD commitment, the final system still eagerly commits to a single syntactic analysis, and in some cases that analysis was wrong (e.g., wrong PP attachment), causing failure for some of the 16 in-scope, positive examples that the final system failed to answer. Clearly deferred commitment can be further extended to explore alternative syntactic analyses. The remaining failures were due to incorrect semantic interpretation of the syntactic analysis, primarily due to poor handling of coordination.

The median, average, and maximum cpu times per question were 0.7, 4.9, and 20.3 seconds respectively.

6 Discussion and Conclusion

Although question interpretation is challenging, we are in the unusual position of having substantial, formal domain (biology) knowledge available. We have illustrated how this knowledge can be exploited to improve question understanding by interleaving interpretation and answering together, allowing the DIRT paraphrase database to be feasibly used and avoiding premature sense commitment. The result is an improved understanding of the original biology questions.

Our work extends previous work (Section 2) on exploring multiple interpretations and exploiting paraphrases, doing so in the context of a task involving formal reasoning. In particular, by interleaving the expansion of possible interpretations with reasoning (that contracts those alternatives), a viable system can be constructed in which the combinatorics are controlled. However, although the system defers WSD and SRL commitment, there are other sources of brittleness – in particular its commitment to a single semantic analysis – that could also benefit from exploration of alternatives, e.g., by using packed representations (Bobrow et al., 2005).

A second limitation of the current approach is that it assumes the (semantics of the) question is a generalized subset of information in (or inferrable from) the KB, i.e., questions are "pure queries" about the KB that do not posit any new information. However some questions, in particular hypotheticals ("X is true. Does Y follow?"), violate this "pure query" assumption by asserting a novel premise (X) that is not in the KB, and hence cannot be disambiguated by searching for the premise X. Although such questions are relatively rare

in biology, they are common in other sciences (e.g., physics). Handling such questions would require extension of this approach, eg by matching a generalized form of the assertion X against the KB to identify how to disambiguate it. Similarly, if we wished to use the system to read new knowledge, as opposed to identify old knowledge, further extensions would be needed, as new knowledge by definition cannot be proved from the KB.

Finally, this work suggests that paraphrase databases such as DIRT offer potential for language understanding in the context of posing formal questions to a reasoning system or database, by bridging gaps that would otherwise have to be hand-engineered, extending their previous use in semi-formal settings such as textual entailment (Bentivogli et al., 2009). Despite noise, the question plus KB help "triangulate" on good paraphrases, and with a suitable user interface to expose their use, this work suggests that there is substantial potential for deploying them in a practical, end-user environment.

Acknowledgements

We are grateful to Vulcan Inc., who funded this work as part of Project Halo.

References

Alshawi H., van Eijck, J. 1989. Logical Forms in the Core Language Engine. Proc ACL, pp25-32.

Bentivogli, L., Dagan, I., Dang, Hoa, Giampiccolo, D., Magnini, B. 2009. The Fifth PASCAL Recognizing Textual Entailment Challenge. In *Proc Text Analysis Conference (TAC'09).*

Bobrow, D. 1964. A Question-Answering System for High School Algebra Word Problems. *AFIPS conference proceedings*, 16: 591-614.

Bobrow, D. G., Condoravdi, Crouch, R. Kaplan, R. Karttunen, L., King, L.T.H., de Paiva, V., Zaenen, A. 2005. A Basic Logic for Textual Inference. In *Proceedings of the AAAI Workshop on Inference for Textual Question Answering*, Pittsburgh, PA.

Chierchia, G. 1993. Questions with Quantifiers. In *Natural Language Semantics* 1, 181-234.

Clark, P. 2009. *A Study of Some "Hard to Formulate" Biology Questions*. Working Note 33, Boeing Technical Report.

Clark, P., Chaw, J., Chaudhri, V., Harrison, P. 2007. Capturing and Answering Questions Posed to a Knowledge-Based System. In *Proc. KCap 2007.*

Clark, P. Harrison, P. 2009. An inference-based approach to textual entailment. In Proc TAC 2009 (Text Analysis conference).

Curtis, J., Matthews, G., Baxter, D. 2005. On the Effective Use of Cyc in a Question-Answering System. *Proc Workshop on Knowledge and Reasoning for Answering Questions*, IJCAI'05, pp 61-70.

Dras, M., Yamamoto, K. (Eds). 2005. *Proc 3rd Internationl Workshop of Paraphrasing*. South Korea.

Harabagiu, S., Moldovan, D., Pasca, M., Mihalcea, R.,et al., 2000. FALCON: Boosting Knowledge for Answer Engines. *Proc TREC'2000 (9th Text Retrieval Conf)*, pp 479-488.

Hobbs, J. Stickel, M., Appelt, D., Martin, P. 1993. Interpretation as Abduction. *Artificial Intelligence* 63 (1-2), pp 69-142.

Kim, D., Barker, K., Porter, B. 2010a. Building an End-to-End Text Reading System based on a Packed Representation. *Proc NAACL-HLT Workshop on Machine Reading*.

Kim, D., Barker, K., Porter, B. 2010b. Improving the Quality of Text Understanding by Delaying Ambiguity Resolution. *Proc COLING 2010*.

Lin, D. and Pantel, P. 2001. Discovery of Inference Rules for Question Answering. *Natural Language Engineering* 7 (4) pp 343-360.

Manning, C., Schutze, H. 1999. *Foundations of Statistical Natural Language Processing.* MA: MIT Press.

Navigli. R. 2009. Word Sense Disambiguation: a Survey. *ACM Computing Surveys,* 41(2), ACM Press, pp. 1-69

Novak, G. 1977. Representations of Knowledge in a Program for Solving Physics Problems, *IJCAI'77*, pp. 286-291

Pinkal, M. 1999. On Semantic Underspecification. In Bunt, H./Muskens, R. (Eds.). *Proceedings of the 2nd International Workshop on Compuational Linguistics (IWCS 2).*

Rinaldi, F., Dowall, J. et al., 2003. Exploiting Paraphrases in a Question Answering System. In *Proc 2003 ACL Workshop on Paraphrasing (IWP 2003).*

Sekine, S., Inui, K. 2007. *Proc ACL-PASCAL Workshop on Textual Entailment and Paraphrasing.*

van Deemter, K., Peters, S. 1996. *Semantic Ambiguity and Underspecification*. CA: CSLI.

Woods, W. 1977. Lunar rocks in natural English: Explorations in natural language question answering. *Fundamental Studies in Computer Science.* A. Zampolli, Ed. North Holland, 521-569.