Recognizing Textual Entailment with Logical Inference

Peter Clark and Phil Harrison
Boeing Phantom Works
The Boeing Company
Seattle, WA 98124
{peter.e.clark,philip.harrison}@boeing.com

Abstract

With the goal of producing explainable entailment decisions, and ultimately having the computer "understand" the sentences it is processing, we have been pursuing a (somewhat) "logical" approach to recognizing entailment. First our system performs semantic interpretation of the sentence pairs. Then, it tries to determine if the (logic for) the H sentence subsumes (i.e., is implied by) some inference-elaborated version of the T sentence, using WordNet (including logical representations of its sense definitions) and the DIRT paraphrase database as its sources of knowledge. For pairs where it can conclude or refute entailment, the system often produces explanations which appear insightful, but also sometimes produces explanations which are clearly erroneous. In this paper we present our system and illustrate its good and bad behaviors. While the good behaviors are encouraging, the primary challenges continue to be: lack of lexical and world knowledge; poor quality of existing knowledge; and limitations of using a deductive style of reasoning with imprecise knowledge. Our best scores were: 56.5% (2-way task) and 48.1% (3-way task)

1. Introduction

Ultimately, an entailment recognition system should not only score well, but also understand and be able to explain why an entailment holds. The latter task is formidable, requiring that the computer form a plausible, coherent, internal model of the "scene" the text describes, requiring vast amounts of lexical and world knowledge. In our work, we are attempting small steps toward this, following a (somewhat) "logical" or "deep" approach to recognizing entailment. First our system (called BLUE, for Boeing's Language Understanding Engine) performs full semantic interpretation of the sentences. Then, it performs inference on those interpretations using WordNet (Fellbaum, 1998), including logical representations of the sense definitions, and the DIRT paraphrase database (Lin and Pantel, 2001), to attempt to relate them. In this paper we present this approach, show examples of both the successes and failures exhibited by the system, and discuss avenues for progressing further along this path.

2. System Description

2.1 Initial Language Processing

We briefly summarize how BLUE converts the initial T and H sentences into logic. Further details are provided in (Clark et al., 2008).

BLUE comprises a parser, logical form (LF) generator, and final logic generator. Parsing is performed using SAPIR, a mature, bottom-up, broad coverage chart parser (Harrison & Maxwell 1986). The parser's cost function is biased by a database of manually and corpus-derived "tuples" (good parse fragments), as well as hand-coded preference rules. During parsing, the system also generates a logical form (LF), a semi-formal structure between a parse and full logic, loosely based on (Schubert and Hwang, 1993). The LF is a simplified and normalized tree structure with logic-type elements, generated by rules parallel to the grammar rules, that contains variables for noun phrases and additional expressions for other sentence constituents. Some disambiguation decisions are performed at

this stage (e.g., structural, part of speech), while others are deferred (e.g., word senses, semantic roles), and there is no explicit quantifier scoping. A simple example of an LF is shown below (items starting with underscores "_" denote variables):

```
;;; LF for "A soldier was killed in a gun battle."
(DECL
    ((VAR _X1 "a" "soldier")
     (VAR _X2 "a" "battle" (NN "gun" "battle")))
   (S (PAST) NIL "kill" _X1 (PP "in" _X2)))
```

The LF is then used to generate ground logical assertions of the form r(x,y), containing Skolem instances, by applying a set of syntactic rewrite rules recursively to it. Verbs are reified as individuals, Davidsonian-style. An example output is:

```
;;; logic for "A soldier was killed in a gun battle."
object(kill01,soldier01),
in(kill01,battle01),
modifier(battle01,gun01).
```

plus predicates associating each Skolem with its corresponding input word and part of speech. At this stage of processing, the predicates are syntactic relations (subject(x,y), object(x,y), modifier(x,y), and all the prepositions, e.g., in(x,y)). Plurality, tense, and aspect are represented using special predicates, asserted for the Skolems to which they apply. Negation is represented by a special assertion that the sentence polarity is negative. Definite coreference is computed by a special module which uses the (logic for the) referring noun phrase as a query on the database of assertions. Another module performs special structural transformations, e.g., when a noun or verb should map to a predicate rather than an individual. Two additional modules perform (currently naive) word sense disambiguation (WSD) and semantic role labelling (SRL). However, for our RTE experiments we have found it more effective to leave senses and roles underspecified, effectively considering all valid senses and roles (for the given lexical features) during reasoning until instantiated by the rules that apply.

2.2 Recognizing Textual Entailment

2.2.1 Subsumption

Given the logic representing the T and H sentences, we treat the core entailment task as determining whether T implies H. Similar to several other RTE systems (e.g., Bobrow et al, 2007), the simplest case of this is if the representation of the H sentence subsumes (is more general than, is thus implied by) that of T. For example, (the logic for) "A person likes a person" subsumes "A man loves a woman". This basic operation is also used to determine if an inference rule's condition is satisfied by a sentence, and thus can be applied.

A set S1 of clauses subsumes another S2 if each clause in S1 subsumes some (different) member of S2. A clause C1 subsumes another C2 if both (for binary predicates) of C1's arguments subsume the corresponding arguments in C2, and C1 and C2's predicates "match". An argument A1 subsumes another A2 if some word sense for A1's associated word is equal or more general (a hypernym of) some word sense of A2's associated word (thus effectively considering all possible word senses for A1 and A2). (We also consider adjectives related by WordNet's "similar" link, e.g., "clean" and "pristine", to be equal.). Two syntactic predicates "match" (i.e., are considered to denote the same semantic relation) according to the following rules:

(i) both are the same
(ii) either is the predicate of(x,y) or modifier(x,y)
(iii) the predicates subject(x,y) and by(x,y) match (for passives)
(iv) the predicates are in a small list of special cases that should match e.g., on(x,y) and onto(x,y).

These rules for matching syntactic roles are clearly an approximation to matching semantic roles, but have performed better in our experiments than attempting to explicitly assign (with error) semantic roles early on and then matching on those.

In addition, in language, ideas can be expressed using different parts of speech (POS) for the same basic notion, e.g., verb or noun as in "The bomb destroyed the shrine" or "The destruction of the shrine by the bomb" (Gurevich et al., 2006). To handle these cross-POS variants, when finding the word senses of a word (above) our system considers all POS, independent of its POS in the original text. Combined with the above predicate-matching rules, this is a simple and powerful way of aligning expressions using different POSs, e.g.:

- "The bomb destroyed the shrine" and "The destruction of the shrine by the bomb" (but not

"The destruction of the bomb by the shrine") are recognized as equivalent.
- "A person attacks with a bomb" and "There is a bomb attack by a person" are recognized as equivalent.
- "There is a wrecked car", "The car was wrecked", and "The car is a wreck" (adjective, verb, and noun forms) are recognized as equivalent.

Although clearly these heuristics can go wrong, they provide a basic mechanism for assessing simple equivalence and subsumption between texts. We give some examples of their good and bad behavior shortly.

2.2.2 Inference

In addition to comparing the (logic for the) T and H sentences directly, our system looks for elaborations of T that are subsumed by H, by applying inference rules to T. A rule is applied if the rule's condition subsumes the T sentence, and if so, asserting the rule's conclusion after binding the shared variables.

We are using two inference rule databases:

- The WordNet glosses, converted into logic (100,000 rules). We are using a combination of our own logical forms (Clark et al., 2008) and those of Extended WordNet (Moldovan and Rus, 2001). For nouns, axioms are of the form:

  isa(N,*noun-sense*) → ... .

  For verbs, the axioms are of the form:

  isa(V,*verb-sense*) [& subject(V,X)]
  [& object(V,Y)] → ... .

  (All possible senses are considered in the elaboration. We discuss the benefits and limitations of this shortly).

- The DIRT inference rule (paraphrase) database (Lin and Pantel, 2001). The database contains 12 million rules, discovered automatically from text, of form (X relation_1 Y) → (X relation_2 Y), where relation is a path in the dependency tree between constitutents X and Y.

Although both databases are quite noisy (as we discuss later), they allow more sophisticated entailments to be both spotted and explained. Figure1 shows an example explanation generated by BLUE.

Pair 215:
T: George Bush, the US President, has arrived in the UK...
H: Bush visits the UK.

Yes! I have general knowledge that:
IF Y **arrives in** X THEN Y **visits** X
Here: X = the UK, Y = Bush
Thus, here:
We are told in T: the Bush arrives in the UK
Thus it follows that: the Bush visits the UK

Hence: **Bush visits the UK.**

Figure 1: BLUE provides friendly, comprehensible explanations for the entailments it finds.

2.2.3 Error Tolerance

Despite the sizes of these two databases, BLUE often misses valid entailments following the algorithm described, often because a single predicate in H does not subsume anything in T (and no inference rules make the connection). To accomodate this, we experimented with 3 variants of our system:

run1: H must subsume the inference-elaborated T
run2: Same, except up to 1 mismatch is allowed, i.e., up to 1 predicate in H is allowed not to subsume the inference-elaborated T for subsumption (entailment) to be recognized
run3: Same except up to 2 mismatches are allowed.

3. Results

The RTE4 test data comprises 1000 T-H pairs, each labeled as either entailed (YES), unknown (UNKNOWN), or contradiction (NO). The 3-way task is to assign these three categories correctly to each pair, the 2-way task merges UNKNOWN and NO into a single category. BLUE was able to generate logic for (both T and H in) 786 of the pairs, allowing inference-based entailment to be attempted for these pairs. The remaining 214 were always classified as UNKNOWN.

Our results are shown in the Table 1. When no mismatches were allowed (run1), the system only made a YES or NO prediction for 62 (6.2%) of the 1000 pairs, rising to 30.9% for 1 mismatch (run2) and 55.7% (run3).

Table 1: Results of BLUE's 3 runs on RTE4. The tables show the number of pairs in each category, with correct answer counts (3-way) in bold. "?" denotes UNKNOWN, "Inference" denotes use of a DIRT or WordNet gloss inference rule.

Method and Prediction → Actual ↓	**Subsumption**		**Inference**		**None**
	YES	NO	YES	NO	?
YES	**12**	2	**23**	0	463
?	5	0	8	0	**337**
NO	2	**2**	5	**3**	138
Score (2-way)	14/23=61%		etc.=67%		51%
Score (3-way)	61%		67%		36%
Overall score	51.5% (2-way), 37.7% (3-way)				

run1: no mismatches allowed. Only in a few (62, 6.2%) cases could entailment be proven/refuted, but with reasonable accuracy for these cases.

Method and Prediction → Actual ↓	**Subsumption**		**Inference**		**None**
	YES	NO	YES	NO	?
YES	**85**	4	**91**	4	316
?	22	2	55	0	**271**
NO	8	**5**	26	**7**	104
Score (2-way)	92/126=73%		etc.=54%		54%
Score (3-way)	90/236=71%		etc.=54%		39%
Overall score	56.5% (2-way), 45.9% (3-way)				

run2: 1 mismatching predicate allowed in subsumption testing. This increased coverage of proving/refuting entailment (30.8%) but with lower accuracy for these cases.

Method and Prediction → Actual ↓	**Subsumption**		**Inference**		**None**
	YES	NO	YES	NO	?
YES	**197**	11	**87**	2	203
?	99	4	61	4	**182**
NO	43	**13**	34	**2**	58
Score (2-way)	214/367=58%		etc.=49%		54%
Score (3-way)	210/367=57%		etc.=47%		41%
Overall score	54.7% (2-way), 48.1% (3-way)				

run3: 2 mismatches allowed, increasing coverage (to 55.7%) but decreasing accuracy for these cases.

4. Analysis

In this Section we illustrate some of the successes of the system, and also some of the cases where errors occurred. In the subsequent discussion we provide some broader discussion.

All the below results are taken from **run1** (no mismatch tolerated). In the pairs below, we use the notation:

H for ENTAILMENT/YES
H? for UNKNOWN
H* for CONTRADICTION/NO

We also abbreviate the examples for presentation purposes.

4.1 Use of basic WordNet

4.1.1 Matching Concepts

Our equality and subsumption testing relies heavily on WordNet's synonym, hypernym, and derivation links. As described earlier, we do not perform word-sense disambiguation during sentence interpretation, but instead, during subsumption testing, search for *any* sense(s) of the word(s) that enable entailment to be concluded. This often leads to success, for example:

155 (BLUE got this right):
T: Apple does not intend to **release** a new iPhone..
H*: Apple will **issue** a new iPhone.

BLUE succeeds by recognizing "release" and "issue" as synonyms. Similarly:

202 (BLUE got this right):
T: ...Chris Martin has **walked out of** an interview...
H: Chris Martin **abandons** an interview.

WordNet's hypernym tree includes that (a sense of) "walk out" is (a sense of) "abandon". However, the heuristic can also go wrong sometimes, for example:

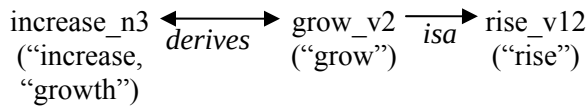
873 (BLUE got this wrong, predicting YES):
T: 433 Soldiers ...**went** missing..on the …range...
H?: 433 Soldiers **died** on the …range.

Here BLUE (undesirably) finds a synonymous sense of "go" ("went") and "die" (namely, pass from physical life), and thus incorrectly concludes T entails H.

BLUE also follows WordNet's derivational links (connecting morphologically related noun and verb senses) to cross part-of-speech boundaries. For example:

27: (BLUE got this wrong, but for other reasons)
T: ...**rising** food prices...
H*: Food prices are **on the increase**.

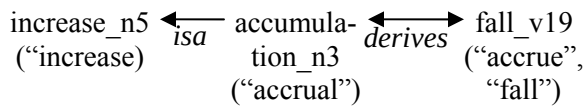
BLUE correctly matches "increase" and "rising" via WordNet's derivational link:



However, this can occasionally also go wrong, for example BLUE (undesirably) matches "increase" and "falling" in:

28 (BLUE got this wrong, predicting YES):
T: ...**falling** food prices.
H*: Food prices are **on the increase**.

via WordNet's links:



where fall_v19 is the sense: "come into the possession of".

4.1.2 Matching Relations

As described earlier, instead of semantic role (predicate) labeling, BLUE heuristically matches syntactic roles (predicates). To a first approximation, the matching rules are: of(x,y) and modifier(x,y) match anything, otherwise the predicates must match exactly. This again has successes, e.g.,:

166 (BLUE got this right):
T: ...convicted **of** child pornography..
H: ...convicted **for** child pornography..

202 (BLUE got this right):
T: ...walked out **of** an interview
H: ..abandons an interview

542 (BLUE got this right):
T: A wax model **of** Amy Winehouse...
H: An Amy Winehouse wax model...

Again, this heuristic can sometimes go wrong. For example, below, BLUE undesirably matches "convicted **of** murder" and "sentenced **to** death" (also using from WordNet that "conviction" and "sentence" are synonyms, and "murder" isa "death"):

865: (BLUE got this right, but for wrong reasons):
T: ...Williams...death row inmate...convicted **of** four murders...
H: ...Williams was sentenced **to** death.

Clearly, allowing of(x,y) and modifier(x,y) to match any predicate is too permissive, although in general this heuristic helps more than it hurts.

4.2 Use of WordNet glosses

Despite having logic for many of the WordNet glosses, they were only of limited help: Of the 39 inference-based entailments in Table 1, only two were successfully due to the glosses. The first is:

177 (BLUE got this right):
T: Corn prices have **hit** new highs...
H: Corn prices **increase**.

using (logic for) WordNet's gloss:

> **score_v1** ("score","hit"): gain points in a game.

and that "gain" isa "increase". This inference is somewhat questionable as the gloss is for the wrong sense of "hit", and as "new highs" has been ignored. If T had said "hit new lows", BLUE would still conclude entailment, this time incorrectly. The second example is:

801 (BLUE got this right):
T: ...Putin has **relieved** from power...Kuroyedov
H: ...Kuroyedov was **fired** by Putin.

using (logic for) WordNet's gloss:

> **take_over_v3** ("take over","relieve"): free someone temporarily from his or her obligations.

and that (a sense of) "free" and "fire" are synonyms.

More commonly, the gloss axioms did not conclude anything relevant to H for two main reasons: First, the gloss definition constitutes only a small part of the knowledge we have about a concept. For example, "marry"(v) is defined as "take in marriage", which omits many of the non-definitional, plausible implications we know about two people marrying (e.g., they probably love each other, live together, are husband and wife, etc.). Second, the quality of the logicized glosses is poor. The gloss English was not written with machine

interpretation in mind, and often includes both gaps and wordy expressions (e.g., "hammer(n): used to deliver an impulsive force by striking"), resulting in logic which is syntactically valid but semantically largely meaningless. In our experiments, it was typically only the very short glosses that were useful.

4.3 Use of DIRT paraphrases

The DIRT paraphrase database (Lin and Pantel, 2001) contains 12 million rules of the form (X relation$_1$ Y) → (X relation$_2$ Y), where relation is a path in the dependency tree/parse between constitutents X and Y. It's vast size was reflected in its use: typically, around 1000 DIRT rules fire on a sentence, compared with around 10 WordNet gloss rules. It contains a mixture of paraphrases, insightful plausible implications, and noise. Informally, about 50% of the DIRT rules seem reasonable. Three examples of success with DIRT are:

(1) 376 (BLUE got this right):
T: ..Billy Connolly is to **star as** …Brodie...
H: …Brodie will be **interpreted by** Billy Connolly.

via WordNet's "portray" isa "interpret" and DIRT's

Y stars as X → Y portrays X;

(2) 870 (BLUE got this right):
T: ...**project of** NASA...and the...Agency.
H: NASA **collaborates with** the...agency.

via WordNet's "project" isa "work" and DIRT's

X works with Y → X collaborates with Y;

(3) 238 (BLUE got this right):
T: ...Ford has **sold** ... Jaguar...**to**...Tata
H: Tata **takes over** Jaguar.

via the DIRT rule:

Y is sold to X → Y is taken over by X

In other cases, an invalid DIRT rule leads to an incorrect conclusion, e.g.,:

167 (BLUE got this wrong, predicting YES):
T: R. Kelly was **acquitted** of child pornography...
H*: R. Kelly was **convicted** for child pornography.

via the (bad) DIRT rule:

Y is acquitted in X → Y is convicted for X

There were several other cases of bad DIRT rules leading to incorrect conclusions. Some of the questionable/invalid DIRT rules which fired were:

Y is criticized by X → Y is endorsed by X
Y is relieved by X → Y is elated by X
Y occurs in X → something dies in X of Y
X is caused of Y → Y is caused by X

4.4 Other special cases

4.4.1 Negation

If negation is encountered, BLUE flags the sentence as having negative polarity (and then inverts the entailment decision accordingly). For example:

155 (BLUE got this right):
T: Apple does **not** intend to release a new iPhone..
H*: Apple will issue a new iPhone.

However, this went wrong in the example below, where the "not" led to the sentence being flagged as having negative polarity even though the "not" was irrelevant to the matching clauses:

27 (BLUE got this wrong, predicting NO):
T: ...rising food prices...if things do **not** ease soon.
H: Food prices are on the increase.

4.4.2 Quantification

BLUE does not take into account plurals or quantification, treating all sentences as being about specific individuals. It thus got the below wrong:

415 (BLUE got this wrong, predicting YES):
T: ...is a scheme for **people who work for the** ...
H*: ..is a scheme for **everyone who works**.

Note that the YES prediction would have been correct if the text had said "a person" (existential quantification) rather than "people"/"everyone" (universal quantification).

4.4.3 Modals

Although BLUE will interpret modal expressions, representing them with second-order logic expressions (where clauses are arguments to other clauses), they were flattened for our RTE application. As a result, BLUE is unable to distinguish between facts and statements *about* (possibly false) facts. For example:

699 (BLUE got this wrong, predicting YES):
T: **Proposals to** extend the Dubai Metro...
H?: Dubai Metro will be expanded.

749 (BLUE got this wrong, predicting YES):
T: ...polls had **predicted that** Donald Tusk...would win the election.
H?: Donald Tusk won the election.

4.5 Overall Results

Despite cases where BLUE makes mistakes, our approach had significant net benefit. For the pairs where BLUE could determine entailment or contradiction, it was correct 65% of the time (run1), 61% (run2), and 54% (run3) on the 3-way task. The most significant limitation was lack of knowledge, and to a lesser extent the deductive style of reasoning. We discuss these further in the next section.

5. Discussion

5.1 Knowledge Limitations

Despite the size of the knowledge sources we are using, when we use a (somewhat) logical approach to entailment (run1), entailment is predicted (or refuted) for only 6.2% (62 pairs), compared with 65% (650 pairs) marked as entailed/contradiction in the official answer key. While a proportion (maybe one third) are due to errors in the initial text interpretation, the majority of failures appear to be due to lack of lexical and world knowledge.

Although BLUE's scores increase when we relax the matching criteria for subsumption (run2 and run3), this is somewhat unsatisfying as there is no longer a completely valid explanation for the predictions. We here discuss some examples of missing knowledge. Also, a detailed discussion of the knowledge requirements for the previous RTE3 is given in (Clark et al., 2007).

The required lexical and world knowledge for RTE4 varies tremendously, from "core" general knowledge to domain-specific facts. For example:

247 (BLUE incorrectly predicted UNKNOWN):
T: ...ice **in** the soil **around** its landing site **on** Mars.
H: Ice...**on** Mars.

requires some core spatial knowledge, here that:

W is in X around Y on Z → W is on Z

Other pairs require more domain-specific knowledge, which we illustrate with two examples. First:

840 (BLUE incorrectly predicted UNKNOWN)
T: A **bus collision**...resulted in...30 **fatalities**...
H: 30 were **killed** in a **road accident**...

Here the system needs to realize that a "bus collision" isa "road accident", a task beyond a simple hypernym check as these are compound nouns (WordNet allows us to conclude "bus collision" isa "accident", but not a "road accident"). Rather, a deep understanding requires knowledge that a bus collision (likely) happened on a road. WordNet has hints of this ("bus: a vehicle...for..transport" and "road: an open way for transportation") but not that a bus is for *road* transportation, and hence we cannot make the complete connection.

Similarly in this example, the system needs to relate "fatality" and "killed", e.g.,

X is a fatality → X was killed

(This is outside the expressive power of DIRT). Again, WordNet gives hints ("fatality: a death resulting from an accident" and "kill: cause to die"), and DIRT tells us that:

X causes Y → X results from Y

but the chain of reasoning is too complex for BLUE to assemble.

As a second example, consider:

83 (BLUE incorrectly predicted UNKNOWN):
T: Clinton...[is]...re-elected...
H: Clinton wins elections.

The challenge here is entailing the "wins" verb in H. This requires knowledge about elections, e.g., that an election is a selection competition, and the winner is the one selected. WordNet again comes close ("elect: select by a vote", "election: a vote to select the winner...") but does not state that the elected one is the winner. This experience of getting "tantalizingly close" using the existing knowledge sources was a common one.

5.2 Reasoning Limitations

A second limitation arises from the use of a "deductive" paradigm with uncertain and noisy knowledge. BLUE recognizes entailment if there is *some* chain of reasoning that connects T with H,

even if there is other peripheral information suggesting otherwise. Although this "tunnel vision" of the system was less of a problem for us than the system's basic lack of knowledge, it clearly will become significant as additional knowledge sources become available.

For example, consider the earlier example:

167 (BLUE got this wrong, predicting YES):
T: Kelly was **acquitted** of child pornography after the star witness ...was discredited...
H*: Kelly was **convicted** for child pornography.

BLUE got this wrong because it found an inference chain using the (bad) DIRT rule:

Y is acquitted in X → Y is convicted for X

However, there is also evidence against this entailment. Most obviously, WordNet has "acquit" and "convict" as antonyms, reinforced by definitions "acquit: pronounce not guilty" and "convict: declare guilty". Also, from world knowledge we know trials with discredited star witnesses rarely result in convictions. As BLUE simply looks for a connecting chain of inferences, this counter-evidence was ignored.

In general, people understand text by creating a "mental model" of the scenario that the text describes, including facts not explicitly mentioned in the text. This is not a deductive process but a search process to find a most coherent model, where coherence is defined with respect to a vast amount of background knowledge and experience. Ultimately, we want our machines to do the same thing.

6. Summary and Conclusion

As a long term goal, we would like the computer to "understand" the sentences it is processing and explain its entailment decisions. To this end, we have been exploring a (somewhat) logical approach to entailment reasoning, using WordNet and DIRT as our main sources of knowledge. For pairs where it can conclude or refute entailment, the resulting system often produces explanations which appear insightful, but also sometimes produces explanations which are clearly erroneous. As described, the three primary limitations were lack of knowledge, poor quality of existing knowledge, and use of a deductive-style approach with uncertain and noisy data. All these reinforce the need for continued research in the acquisition and use of semantic knowledge for language-oriented and other tasks in AI.

References:

Bobrow, D., Condoravdi, C., Crouch, R., de Paiva, V., Karttunen, L., King, T., Nairn, r., Price, L., Zaenen, A. 2007. Precision-Focused Textual Inference. In: *Proc. 2007 ACL-PASCAL Workshop of Textual Entailment and Paraphrasing.*

Clark, P., Murray, W., Thompson, J., Harrison, P., Hobbs, J., Fellbaum, C. 2007. On the Role of Lexical and World Knowledge in RTE3 In: *Proc. 2007 ACL-PASCAL Workshop of Textual Entailment and Paraphrasing.*

Clark, P., Murray, W., Thompson, J., Harrison, P., Hobbs, J., Fellbaum, C. 2008. Augmenting WordNet for Deep Understanding of Text. in *Semantics in Text Processing (Proceedings of STEP 2008*), Ed: J. Bos, R. Delmonte.

Clark, P., Harrison, P. 2008. "Boeing's NLP System and the Challenges of Semantic Representation", in *Semantics in Text Processing (Proceedings of STEP 2008*), Ed: J. Bos, R. Delmonte.

Gurevich, O., Crouch, R., King, T., de Paiva, V. 2006. "Deverbal Nouns in Knowledge Representation". Proc FLAIRS'06.

Fellbaum, C. 1998. "WordNet: An Electronic Lexical Database." Cambridge, MA: MIT Press.

Harrison, P., and Maxwell, M. 1986. "A New Implementation of GPSG", Proc. 6th Canadian Conf on AI (CSCSI'86), pp78-83.

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

Moldovan, D., and Rus, V. 2001. "Explaining Answers with Extended WordNet", in Proc. ACL.

Schubert, L., and Hwang, C. 1993. "Episodic Logic: A Situational Logic for NLP". in "Situation Theory and Its Applications", pp303-337.