An Inference-Based Approach to Recognizing Entailment

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

Abstract

For this year's RTE challenge we have continued to pursue a (somewhat) "logical" approach to recognizing entailment, in which our system, called BLUE (Boeing Language Understanding Engine) first creates a logic-based representation of a text T and then performs simple inference (using WordNet and the DIRT inference rule database) to try and infer a hypothesis H. The overall system can be viewed as comprising of three main elements: parsing, WordNet, and DIRT, built on top of a simple baseline of bag-of-words comparison. Ablation studies suggest that WordNet substantially improves the accuracy scores, while, somewhat suprisingly, parsing and DIRT only marginally improve the accuracy scores. We illustrate and discuss these results. Overall, BLUE's reasoning is sometimes insightful but sometimes nonsensical, the primary challenges being noise in the knowledge sources, lack of world knowledge, and the difficulty of accurate syntactic and semantic analysis. Despite these challenges, we argue that forming semantic representations is a necessary first step towards the larger goal of machine reading, and worthy of further exploration. Our best scores were 61.5% (2 way), 54.7% (3 way), and F=0.29 (Search Pilot).

1. Introduction

Ultimately we would like machines to be able to "read" and fully understand text, forming an internal, semantic representation of its contents that supports inference, explanation, and question-answering. Towards that end, despite its formidability, we continue to pursue a (somewhat) "logical" approach to recognizing textual entailment, in which our system (BLUE, Boeing Language Understanding Engine) constructs and performs simple inference with a logic-based representation of the text. When successful, this approach can infer entailment with a coherent and insightful line of reasoning. However, it can also be unsuccessful, with either an incoherent line of reasoning or (more commonly) no result at all, the primary challenges being being noise in the data sources, lack of world knowledge, and the difficulty of accurate syntactic and semantic analysis.

In this paper we first describe BLUE, its performance on the RTE5 Main Task, and the results of ablation studies. The studies showed WordNet helping substantially, while parsing and DIRT produced only marginal improvements, and we discuss reasons for these results. We also describe use of BLUE on the RTE5 Search Pilot, and how it was modified to account for the greater use of context in that task. Despite the challenges, we argue that the current approach is a small step in the right direction towards both improved RTE performance, and the wider goal of machine reading.

2. System Description

The basic operation of our entailment system, BLUE (Boeing's Language Understanding Engine) is to convert the T and H sentences into a logic-based representation, and then search to see if T implies (or contradicts) H using inference rules from WordNet and the DIRT database.

The system has progressed in two ways since 2008. First, as well as doing inference with the logic representation derived from the parse tree, we have added a second module that performs word-level inference with just the bags of words in T and H, i.e., ignoring syntactic structure. Second, the basic engineering of the language engine BLUE has substantially improved, enabling it to interpret a wider variety of grammatical constructs than in 2008.

The overall system now consists of two entailment modules in a pipeline (Figure 1): The first generates and compares a logical representation of the T and H texts to try and conclude entailment or contradiction. If either can be concluded, the module exits with that conclusion. If not (i.e., the "unknown" cases), the second module performs a similar comparison but using just the bags of words in the T and H texts, i.e., ignoring the syntactic (parse) structure of the texts. WordNet and DIRT can be used in both modules, as we describe shortly. We now describe the logic module and bag-of-words module in turn.

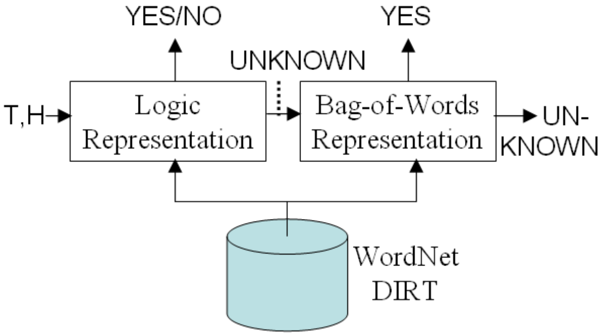

Figure 1: BLUE consists of two pipelined modules, performing simple inference with structured and word-based representations of the text.

2.1 The Logic Module

2.1.1 Initial Language Processing

We briefly summarize how BLUE converts the initial T and H sentences into logic. Further details are provided in (Clark and Harrison, 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). 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. Pronoun and definite reference resolution is performed 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 unspecified, effectively considering all valid senses and roles (for the given lexical features) during reasoning until instantiated by the rules that apply.

2.1.2 Recognizing Textual Entailment

a. 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). 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.

b. Use of WordNet

BLUE makes use of WordNet to recognize the equivalence (synonym) and subsumption (hypernym) relationships between (senses of) words-during the subsumption tests just described. When comparing words, BLUE considers all possible senses. It also ignores the original part of speech of the words, thus effectively considering all parts of speech so that cross-part-of-speech equivalences such as "run"(n) and "run"(v) are recognized.

In addition to synonyms and hypernyms, BLUE uses WordNet's "similar" (SIM), "pertains" (PER), and "derivational" (DER) links to recognized equivalence. The "similar" links connect approximately equivalent (senses of) adjectives, e.g.,

speedy#s2 ←similar-to→ fast#a1

The "pertains" links connect approximately equivalent (senses of) nouns, adjectives, and adverbs, e.g.,

rapidly#r1 ←pertains-to→ quick#a1

The "derivational" links connect equivalent (senses of) nouns and verbs, e.g.,

destroy#v1 ←derives→ destruction#n1

thus enabling verbs and nominalizations to be related (Gurevich et al., 2006). These additional WordNet relations enable a substantially larger number of word equivalences to be correctly recognized than just using the synonym relations.

c. Use of DIRT Inference Rules

In addition to comparing the (logic for the) T and H sentences directly, BLUE 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, the rule's conclusion is asserted after binding the shared variables.

Our source of inference rules is the DIRT inference rule (paraphrase) database (Lin and Pantel, 2001, Pantel et al, 2007). The database contains 12 million rules, discovered automatically from text, of form:

$$(X\ relation_1\ Y) \rightarrow (X\ relation_2\ Y)$$

where *relation* is a path in the dependency tree between constitutents X and Y. Although the database is quite noisy, it allows more sophisticated entailments to be both spotted and explained.

We found that the verb rules (e.g., IF X loves Y THEN …) were substantially more reliable than the noun rules (e.g., IF X part of Y THEN …), and as a result only use the verb paraphrases in BLUE.

2.1.3 Error Tolerance

Despite the sizes of WordNet and DIRT, 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 allow up to 1 mismatch during subsumption testing, i.e., up to 1 predicate in H is allowed not to subsume the inference-elaborated T for subsumption (entailment) to be recognized. We also experimented with allowing more (2) and fewer (0) mismatches. Allowing more mismatches results in greater coverage but less accuracy and worse[1] explanations in the logic module, with the overall system accuracy remaining largely unchanged. 1 mismatch seemed to provide the best balance of coverage and explanation quality.

2.2 The Bag-Of-Words Module

The bag-of-words module is used if the logic module is unable to conclude or refute entailment. It performs similar inference-based comparisons to conclude entailment, but between the bags of words rather than logic for the T and H sentences, thus ignoring syntactic structure. Each bag is the collection of the (root forms of the) nouns, verbs, adjectives, and adverbs in a sentence (thus prepositions, determiners, etc. are ignored), and also ignoring the verb "be". To compute subsumption with bags, BLUE searches for some pairing of H-words with T-words such that each H-word subsumes a T-word. Note that a word cannot be used twice in the pairings (unless it occurs twice in the bag), and thus subsumption involves a search to find the best pairing. As the bags are small an exhaustive search is straightforward.

a. Use of WordNet

WordNet is used to compute equivalence and subsumption between words in this module in the same way as for the logic module.

b. Use of DIRT Inference Rules

A large number of the DIRT paraphrases are of the form:

IF X *verb* Y THEN X *verb'* Y

Because the dependency paths in the condition and action are the same here, we can infer a word-level substitution inference that *verb* → *verb'*. In many cases these duplicate WordNet's synonym and hypernym relations, but in many cases they denote new inferential relationships outside WordNet, such as "kiss" → "love", "meet" → "visit", and "market"(v) → "sell". BLUE uses these DIRT-derived inferential relationships when computing subsumption between words in the bags.

3. Results – Main Task

We ran three configurations of BLUE: The bag-of-words module alone; the logic module alone; and the logic plus bag-of-words module in the pipeline (i.e., the full system). The results are shown in Table 1.

	2-Way	**3-way**
Bag-of-words module only	60.0	52.8
Logic module only	56.7	46.3
Logic + bag-of-words	**61.5**	**54.7**

Table 1: BLUE's scores (% correct) on the RTE5 test set. The best results were obtained when using the logic + bag-of-words pipeline.

The best results were obtained with the pipeline, using first the logic-based and then word-based entailment. The lower score for the logic module alone is primarily due to the low number of cases for which it could infer (or refute) entailment: Of the 600 pairs, it inferred (or refuted) entailment for 176 (29%) of them with a relatively high (63.6%) accuracy, but the remaining 424 were then labelled "unknown". Thus adding the bag-of-words module to handle these unknown cases substantially improved the overall score. It is also interesting that the bag-of-words module alone (but still using WordNet and DIRT for inference within it) is only slightly worse than the pipeline, suggesting that extracting syntactic structure provides only small additional discriminatory power. We discuss this further in the ablation studies later in Section 4.3.

4. Analysis

In this Section we illustrate and discuss some of the successes and failures of BLUE. All the below results are taken from the full pipelined (logic + bag-of-words) system described earlier. We use the notation:

H for ENTAILMENT/YES
H* for CONTRADICTION/NO

and also abbreviate the examples for presentation purposes.

[1] i.e., the line of reasoning is nonsensical

4.1. Use of WordNet

WordNet was an important source of information relating words together. For example, below:

191 (BLUE got this right):
T: Ernie Barnes... was an offensive lineman...
H: Ernie Barnes was an athlete.

BLUE got this right as WordNet states that (a sense of) "athlete" subsumes (is a hypernym of) a sense of "lineman". Similarly:

467 (BLUE got this right):
T: Katrina...made landfall in...Florida...
H: Katrina hit Florida.

as a sense of "hit" (namely, reach a destination, "We hit Detroit") subsumes a sense of "make" (namely "reach", "We made it to the plane").

WordNet's "similar", "pertains", and "derivational" links were also useful, for example:

303 (BLUE got this right):
T: ...Japanese capital of Tokyo...
H: Tokyo is the capital of Japan.

Here BLUE correctly related "Japanese" and "Japan" because Japanese#a1 pertains to Japan#n2 in WordNet. However, here:

281 (BLUE got this wrong, predicting YES):
T: Clarkson died...
H*: Actress Lana Clarkson killed...

BLUE incorrectly concluded "kill" subsumes "die" because the derivational (DER) nominalizations of these words subsume each other (death#n7 isa killing#n2). In this case, BLUE's heuristic of considering all possible senses has caused the problem, as this sense of death#n7 ("the act of killing") is incorrect for this text.

4.2 Use of DIRT Inference Rules

The 12 million DIRT inference rules are a mixture of equivalences, insightful plausible implications, and noise. Informally, about 50% of the DIRT rules seem reasonable. Some successful and unsuccessful examples using DIRT are:

333: (BLUE got this right)
T: ...an attempted hijacking of a Norwegian tanker...by Somali pirates...
H: Somali pirates attacked a Norwegian tanker.

BLUE correctly inferred entailment via the DIRT rule "IF X hijacks Y THEN Y is attacked by X". Similarly:

26: (BLUE got this right)
T: The U.S. holds about 240 men at the U.S. base in Cuba...
H: About 240 people are detained in Guantanamo.

BLUE correctly inferred entailment via the DIRT rule "IF Y is held by X THEN Y is detained by X", plus WordNet's assertion that "person" subsumes "man". (BLUE did not equate "Guantanamo" with "U.S. base in Cuba", but one mismatch is tolerated during reasoning).

Two example failures with DIRT are:

30: (BLUE got this wrong, predicting YES)
T: A man has hijacked a passenger plane in the Jamaican resort of Montego Bay...
H*: A plane crashed in the Jamaican resort of Montego Bay.

via the (vaguely plausible) DIRT rule "IF Y is hijacked in X THEN Y crashes in X", and

407: (BLUE got this wrong, predicting YES)
T: Venus Williams triumphed over …Bartoli…
H*: Venus Williams was defeated by…Bartoli…

via the (non-sensical) DIRT rule "IF Y wins over X THEN X defeats Y" (and WordNet's "triumph" isa "win").

In general, using DIRT results in a mixture of coherent and incoherent rules and reasoning. A random sample of DIRT rules used by BLUE in the RTE5 challenge illustrate the mixture of good, bad, and simply nonsensical:

IF X is canceled on Y THEN X is won on Y
IF Y takes a X post THEN X elects a Y leader
IF Y is made amid X THEN Y comes of X
IF X is done of Y THEN X does of Y
IF Y causes a death of X THEN Y causes X's death
IF Y wins by X THEN Y wins a election by X
IF someone expects by X in Y THEN Y develops a product in X
If Y sell's X's business THEN Y holds X's tongue
IF X orders on Y THEN X tells a President on Y

4.3 Ablation Studies

BLUE can be considered to use three sources of knowledge to infer entailment: the syntactic structure (parse) of the sentences, WordNet, and DIRT. We ran three ablation studies on the full pipelined system to investigate the relative contribution of each of these sources. To ablate the syntactic structure, we bypass the logic module and just use WordNet and DIRT for inferring subsumption with the bag-of-words. The results are shown in Table 2.

Configuration:	RTE5 Dev	RTE5 Test
BLUE (full)	63.8	61.5
- without parse	63.5	60.0
- without DIRT	63.3	62.7
- without WordNet	57.8	57.5

Contribution:	RTE5 Dev	RTE5 Test
parse	+0.3	+1.5
DIRT	+0.5	-1.2
WordNet	+6.0	+4.0

Table 2: Ablation studies show WordNet helping substantially, and parsing and DIRT marginally. Figures are % correct, 2-way test, on the RTE5 development & test sets respectively. Contribution is the difference between ablated and full BLUE.

The main observation from these results is that WordNet is substantially improving the score, while, somewhat surprisingly, parsing and DIRT are barely improving the accuracy scores (and in one case hurting the score).

Concerning DIRT, there appear to be two factors contributing to its limited utility. First, the database is noisy. Despite the presence of many insightful rules (it is remarkable that machine learning has been so successful at acquiring them), about half the rules appear to be invalid or nonsensical. Second, and perhaps more significantly, the DIRT rules only solve a minority of cases. Of the 600 cases, only 86 (RTE5 DEV) and 64 (RTE TEST) used DIRT rules to establish entailment. If half of these rules are bad, then that means only approximately 30-40 (≈ 5%) entailments are using good DIRT rules, and compounded with other possible errors (parse, word sense, etc.) is barely enough to make a significant performance difference.

The main reason DIRT only potentially solves a small number of cases is simply that most RTE pairs require substantially more than a simple inference rule to solve -the RTE test is hard. For example:

157 (BLUE got this wrong, saying UNKNOWN)
T: Slumdog Millionaire director Danny Boyle....The...filmmaker told....
H: The movie "Slumdog Millionaire" has been directed by Danny Boyle.

Here the word "movie" in H causes BLUE's failure. It is only because we know from general knowledge that Slumdog Millionaire *is* a movie (not a company or a play, say) that a person concludes entailment, but BLUE does not have this knowledge. (Consider, for example, that if we replace "Slumdog Millionaire" with "Hollywood" in both T and H then H would not be entailed). Similarly:

499: (BLUE got this wrong, saying UNKNOWN)
T: ...the party, which is backed by Putin, officially nominated Medvedev for presidency...
H: Vladimir Putin supports Medvedev.

Here entailment can be inferred with the plausible inference that if X backs Y, and Y nominates Z, then X supports Z. However, this is beyond DIRT's expressive power.

It is also interesting and somewhat surprising that extracting syntactic structure has, at least in BLUE, only a small effect on accuracy. One reason for this is that extracting syntactic (hence semantic) structure is very error-prone. However, another reason seems to be that the (usually implicit) semantic relationships often remain the same between T and H (even if H is not entailed), reducing the discriminatory power of the syntactic structure. In other words, if the T and H sentences make sense, are coherent with general knowledge, and are topically similar, then this vastly constrains the allowable semantic relationships that can connect the concepts, hence reducing the value of comparing them. Of course, one can *in principle* change the semantic relationships between T and H (e.g., T: "Danny directed the movie", H: "The movie directed Danny") but in practice such rearrangements are usually nonsensical, off topic, or inconsistent with general knowledge, and thus do not occur often in the texts. This is not a general property of language, but it is a general property of naturally occurring texts in the world - we might call it "semantic continuity" - including those texts used in the RTE challenges. Burchardt et al. made similar observations about the limited utility of semantic analysis, although attributed it more to technical challenges (in particular semantic role filler matching) rather than "semantic continuity".

5. The RTE Search Pilot

5.1 Representing Context

In addition to the Main Task, RTE5 included a “Search Pilot” task to identify the sentence(s) T(s) in a newswire article with a headline HEADLINE that entails a hypothesis H. Nine collections of articles and hypotheses were used, each on a different topic. This task differs somewhat from the Main Task in that there is a greater need to take context into account. For example, all the articles in topic D0908-A talk about Nepal, and thus a sentence like:

T: The EU's Luxembourg presidency called for a "speedy" return to multi-party democracy.

implicitly refers to democracy in Nepal, and thus can be considered to entail the hypothesis:

H142: There has been an international call for a return to democracy in Nepal.

even though "Nepal" is not explicitly mentioned or indirectly referred to in the sentence T. (The sentence in isolation does not entail H142.) Some of the Main Task examples also require using the context of the surrounding sentences, but generally to a lesser degree.

To handle this we use a simple approach, namely taking the article's headline as the context for the article's sentences. Specifically, rather than requiring every part of a hypothesis H to be entailed by a sentence T, we only require those parts not mentioned in the headline to be entailed (we assume the other parts are true based on the headline). For logical entailment, a "part" is a single *relation*(*object,object'*) assertion, and it is considered "mentioned" in the headline (thus assumed true) if the headline contains at least one of the words denoting *relation*, *object*, or *object’*. For example, given BLUE's interpretation of H142:

```
; H142: There has been an international call
for a return to democracy in Nepal.
modifier(call01,international01),
for(call01,return01),
to(return01,democracy01),
in(democracy01,nepal01).
```

and an article with the headline:

HEADLINE: EU slams Nepalese king's dismissal of government.

then the last clause, in(democracy01,nepal01), would be assumed because (an inflection of) "Nepal" is mentioned in the headline. As a result, only the first 3 clauses are used to assess entailment between T and H, corresponding to a “reduced hypothesis”:

H142': There has been an international call for a return to democracy.

For bag-of-words entailment, a hypothesis word is assumed if that word is also mentioned in the headline, thus again assuming the headline as context. For example, for H142 above the hypothesis bag of words is:

```
; H142: There has been an international call for a
return to democracy in Nepal.
{"international" "call" "return" "democracy"
"Nepal"}
```

but as (an inflection of) “Nepal” is mentioned in the headline, it is assumed and entailment is judged using the reduced bag:

```
{"international" "call" "return" "democracy"}
```

In general this simple heuristic works well, but it becomes questionable in cases where the headline and hypothesis are similar (in which case almost all of the hypothesis is assumed). For example:

HEADLINE: Prince Charles to Marry Camilla Parker Bowles
H26: Prince Charles was married to Princess Diana

here "Prince Charles" and "married" are assumed, and so the reduced hypothesis becomes simply (the logic/words for) "Princess Diana", i.e., any sentence (in the article with that headline) mentioning Princess Diana is considered entailing H26. In fact, in that article most sentences mentioning Princess Diana do, in fact, refer to her marriage to Prince Charles, and so in this case at least the method is still effective. In the limiting case, though, where the hypothesis and headline are identical, all sentences would be (undesirably) treated as entailing the (empty) hypothesis. To guard against this, we require that there is at least one clause/word in the reduced hypothesis. If there is not, then we perform search with the original hypothesis instead.

5.2 Results

We ran three versions of BLUE:

bag0: Just the bag-of-words module. If the (reduced) H bag subsumes the T bag, then entailment is concluded.

bag1: Just the bag-of-words module, but allowing up to 1 mismatch, i.e., if all but one word in the H bag subsumes the T bag, then entailment is concluded.

logic1: Just the logic module, allowing up to 1 mismatch, i.e., if all but one clause in the H representation subsumes the T representation, then entailment is concluded.

In all three cases, WordNet and DIRT were used to determine entailment between words/clauses. We did not experiment with pipelining the two modules, as we did for the Main Task, although the Main Task results suggests this might be the overall best combination. The results (overall microaverages) are shown in Table 3.

	Precision	**Recall**	**F-Measure**
BLUE-bag0	61.5	15.0	0.24
BLUE-bag1	**33.4**	**25.1**	**0.29**
BLUE-logic1	36.4	9.4	0.15

Table 3: Performance (percentages) of three configurations of BLUE on the RTE Search Pilot.

As would be expected, allowing a mismatch reduces precision but increases recall, with the best tradeoff (F-Measure) being the bag-of-words module allowing 1 mismatch, using WordNet and DIRT to compute subsumption.

5.3 Analysis

5.3.1 Successes and Failures

Sometimes use of WordNet and DIRT lead to some good lines of reasoning in this task, e.g.,

(BLUE got this right)
T: ...France reached a compromise with Israel to bury Arafat in Ramallah...
H97: France and Israel seemed to agree on burial in Ramallah.

was found to be entailed using WordNet's "agree" isa "compromise", and "bury" and "burial" are equivalent according to WordNet's derivational (DER) links. Similarly,

(BLUE got this right)
HEADLINE: Nepal's king may seek…
T: The United States has given Katmandu ...aid...
H147: The United States provided aid to Nepal.

H147 was found entailed by T using WordNet's "give" isa "provide", and assuming "Nepal" because of the headline.

A common cause of false negatives (missed entailments) was when there was implicit knowledge in the texts that was explicit in the hypothesis, and which could not be inferred from the HEADLINE context. For example:

(BLUE got this wrong, predicting NO)
T: If no cardinal wins...the cardinals will pause...as John Paul ordered.
H74: New rules…were introduced by John Paul II.

Here, BLUE could not infer entailment of the hypothesis phrase "new rules" as it is not present (either explicitly or inferentially) in the text. In addition, "order" (T) and "introduce" (H74) are not related in either WordNet or DIRT; in fact, the semantic relationship between these two words is quite complex and their equivalence is context-dependent, beyond that which simple hypernym or paraphrase relationships can capture.

A common cause of false positives (incorrectly found entailments) was when WordNet or DIRT incorrectly related two words. This was particularly true for general words, for example, in:

H19: Titan has an atmosphere.

the word "have" subsumes many words in the T sentences due to its polysemy, including: "enter", "shape", "time", "say", "discover", "return", "land", and "take", often leading to false positives. Similarly:

(BLUE got this wrong, predicting YES)
T: Black smoke signaled that they had voted...
H82*: A signal of white smoke indicates election...

BLUE (undesirably) equated "black" and "white" via WordNet's similar adjective (SIM) links:

black#s14 -similar→ clad_a1 ←similar- white#s10

where the senses are defined as:

black_s14: dressed in black, e.g., "black friars"
white_s10: dressed in white, e.g., "white nuns"

The problem here appears to be a misuse of the "similar" link in WordNet. Finally, and quite humerously, BLUE had a large number of false positives in articles about the pope and his cardinals because "cardinal" can also mean a number (cardinality). As a result, BLUE equated "cardinal" with any number found in the texts, e.g.,:

(BLUE got this wrong, predicting YES)
T: ...smoke from burned ballot papers...could be seen at...around 7 p.m.
H76*: 115 cardinals participated in the pope's election.

via, among other things, the (undesirable) connection "7" isa "cardinal" (!).

The DIRT paraphrases similarly enabled some good entailments, e.g.,

(BLUE got this right, predicting YES)
T: Bobby Fischer...awaiting deportation...
H28: Bobby Fischer faced deportation…

via the DIRT rule "IF X awaits Y THEN X faces Y", and similarly:

(BLUE got this right, predicting YES)
T: ...Charles divorced Diana...
H26: Prince Charles was married to Princess Diana

via the DIRT rule "IF X divorces Y THEN X marries Y" (though BLUE is not doing any temporal reasoning here). As in the Main Task, though, the DIRT rules can result in bad entailments also.

Finally, failing to account for negation, modals, and hypotheticals caused some failures. For example:

(BLUE got this wrong, predicting YES)
T: ...even if Iceland offered Fischer citizenship...
H29*: Iceland granted Bobby Fisher citizenship.

the phrase "even if" turns T into a hypothetical, but BLUE takes T as an actual. (the DIRT paraphase "IF X offers Y THEN X grants Y" completes the connection). Similar problems occurred for the phrases "if convicted" and the negation word "against" in "voted against", which reverses the polarity of the DIRT-based equality "voted" → "granted".

5.3.2 The Precision/Recall Tradeoff

Interestingly, the precision/recall values vary considerably for different hypotheses. BLUE finds an entailment if the concepts in H subsume those in T, and thus in the extremes:

(1) If recall is low, then BLUE has missed some texts T entailing H, suggesting that the hypothesis H can be stated using other concepts/phrases than those in or inferable from H (hence BLUE fails to find them)
(2) If precision is low, then BLUE has found some texts T *not* entailing H, suggesting that the concepts in the hypothesis can be used to state other things besides H (which BLUE has incorrectly taken as entailing H).

We can in fact see these extremes in the results. An example of low recall is:

H8: Prince Charles will marry Camilla Parker Bowles.

Here, BLUE got high precision (0.8571) but low recall (0.1765), i.e., BLUE missed a lot of cases. In other words, for this hypothesis, there are likely many other ways of saying the same thing. And indeed this is what we see looking at the "gold standard" entailing sentences:

- the names are often not fully mentioned (we see "Charles", "Camilla", "the couple", "I", "her", "they")
- many indirect references to "marry ("wed", "wedding", "important step", "engagement", "wife", "proposed to", "will become the Princess Consort")

An example of low precision is:

H147: The United States provided aid to Nepal.

Here BLUE got low precision (0.2174) but high recall (0.8333), i.e., BLUE found too many cases. In other words, for this hypothesis, the words can be used to say something different. In this case the low precision arose because BLUE allows 1 mismatch (thus "United States" can be ignored), and many sentences describe *Japan* providing aid to

Nepal. Thus for this hypothesis, there is an alternative hypothesis which is similar to H147, coherent with world knowledge, within the topical context of the articles, and actually occurred - an unusual combination, and relatively rare in this Pilot task.

6. Summary and Conclusion

In both the Main Task and Search Pilot, BLUE was able to find entailments with sometimes insightful reasoning, and sometimes nonsensical reasoning. Ablation studies showed WordNet significantly helping, and DIRT and parsing marginally helping. BLUE's performance was above the median for the RTE5 Main Task, but still remains limited by errors in the knowledge sources and the parsing/semantic analysis process, combined with lack of knowledge to bridge the often wide semantic gap between the T and H sentences. Although parsing and extracting semantic structure provided only a small benefit in terms of accuracy, it does result in more coherent and valid explanations for entailment (compared with bag-of-words), and ultimately is a first step towards constructing a richer model of the texts.

What would it take to excel at the RTE challenge? While substantial improvement can likely be achieved by improving the engineering and tuning of BLUE, we believe that ultimately the computer needs to better "understand" the texts, i.e., form a single, coherent, inference-supporting representation of their meaning in order to perform well - a goal we are still a long way from achieving. In particular, BLUE's search for *some* reasoning path from T to H without trying to form a deeper model is a crude approach, because without a deeper model there is little basis for distinguishing good paths from bad, or spotting glaring contradictions between the chain of reasoning and other "obvious" implications of the text. Instead, one would like the computer to build a richer model of the text, somewhat independent of the hypotheses to be tested, and likely heavily guided by prior expectations about the world, and then use that model to constrain the entailment paths considered when performing textual entailment. We believe that this direction is one which will ultimately be fruitful both for RTE and in the larger quest for building machines that can read in the future.

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.*

Burchardt, A., Pennacchiotti, M., Thater, S., Pinkal, M. "Assessing the Impact of Frame Semantics on Textual Entailment", *Natural Language Engineering* 1 (1) pp1-25, 2009.

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.

Pantel, P., Bhagat, R., Coppola, B., Chklovski, T., Hovy, E. 2007. ISP: Learning Inferential Selectional Preferences. In Human Language Technologies, NAACL HLT 2007.

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