BLUE-Lite: a Knowledge-Based Lexical Entailment System for RTE6

Peter Clark[1] and Phil Harrison
Boeing Research & Technology
The Boeing Company
Seattle, WA 98124

Abstract

In this paper we present our RTE6 system, BLUE-Lite, and the results of experiments with it. Unlike our earlier RTE5 system, called BLUE, BLUE-Lite uses only a lexical ("bag of words") representation of the sentences. To compare lexical items, BLUE-Lite exploits linguistic and world knowledge drawn from WordNet and the DIRT paraphrase database. To take context into account, BLUE-Lite also looks in the preceding sentence (with reduced confidence) if an H word does not match T. In addition, the entailment theshold is varied between topics to account for the fact that some topics are harder to find entailments in than others. Our results show that WordNet, DIRT, and these two techniques all improved performance (producing an overall F=0.44), and also that a relatively simple baseline ("match all but one") without any of these techniques achieved a surprisingly high score (F=0.40). Finally, we discuss the role of structural information, why it is challenging to yield advantage from it (in particular in this year's challenge), but why ultimately it must be taken into account for further improvements in performance.

1. Introduction

Recognizing Textual Entailment (RTE) is the general task of deciding if one text H is entailed by another text T. In this year's competition, RTE6, both T and H were single sentences, drawn from (different) newpaper articles about the same topic. Each H sentence was manually shortened and simplified from the original sentence by the competition organizers. The candidate T sentences were those that seemed superficially similar to that simplified H sentence (as determined by the Lucene search engine, taking H as a query) although in practice only about 5%-10% of the candidates Ts actually entailed H, as judged by human evaluators. Entailment decisions required taking T's textual context (surrounding sentences) into account, e .g., to determine the identity of references in T.

Our RTE6 system, called BLUE-Lite, is a derivative of our RTE5 system, BLUE (Clark and Harrison, 2009), and is characterized by the following features:

1. The core of the system performs a lexical ("bag of words") comparison between T and H, where: the bag includes not just simple words but also multiwords present in WordNet (we will henceforth simply say "word", but bear in mind this includes any multiword in WordNet); stop words and some other word categories are ignored; and, proper names are compared in a special way.
2. WordNet and DIRT are used to determine lexical entailment between a word in T and a word in H. Our approach could thus be described as "knowledge-based lexical entailment".
3. To account for coreference and context, if a single H word does not match T then BLUE-Lite also looks (with reduced confidence) in the sentence preceding T.
4. Some topics were intrinsically harder to find entailments in than others. To account for this, the threshold for concluding entailment was gradually relaxed to ensure that a minimum

[1] Current address: Vulcan Inc., 505 Fifth Ave S, Seattle, WA 98104. peterc@vulcan.com

number of entailments were found in each topic.

Most strikingly, unlike BLUE, BLUE-Lite makes no use of structural (parse-based) information in entailment decisions. In RTE5, BLUE concluded entailment if either a structural (syntactic) comparison of T and H suceeded (tolerating up to 1 mismatch, or a knowledge-based lexical (bag-of-words) match succeeded. In this year's task, BLUE was finding only a small number of entailments using this approach, so we modified the bag-of-words comparison to also tolerate 1 mismatch. This significantly improving the F-score, but also removed the need for the structural comparison: If the bag-of-words comparison succeeded with 1 mismatch, then it necessarily follows that the structural comparison would also succeed with 1 mismatch (as it is a stricter comparison), obviating the need for the structural comparison in the first place. Informal experiments with allowing 2 mismatches in the structural comparison produced poor results, and so for this year we only used the bag-of-words module, as the pipeline architecture of BLUE made the structural comparison redundant. Although there are alternative ways of combining structural and lexical information (e.g., weighted voting) in which structural information would not be redundant, we have not explored those, and thus there is still potential for further improvement by re-introducing structural analysis into future versions of the system.

In the rest of the paper we first describe BLUE-Lite in detail. We then summarize and give examples of its performance in RTE6, including a characterizations of the various ways in which it fails on some entailments. Finally, we discuss why a purely lexical approach performs relatively well, but why ultimately structural analysis needs to be reintroduced to obtain further performance improvements.

2. System Description

BLUE-Lite's four main characteristics are:

1. selectively generating and comparing a bag of words for T and H,
2. use of lexical and world knowledge from WordNet and DIRT for these comparisons,
3. use of the sentence preceding T to account for context and coreference, and
4. varying the entailment threshold depending on topic.

We describe each of these in turn.

2.1 Lexical Analysis of T and H

To compare T and H sentences, BLUE-Lite first converts each sentence into a bag of "words", where a "word" can be a normal word or a multiword (compound noun, phrase, etc.) present in WordNet's vocabulary. Thus, for example, "Irish Republican Army" is treated as a single (compound) word as that phrase is a member of a WordNet synset (IRA#n1). To generate the bag, the sentence is first parsed to identify words (including multiwords) and parts of speech in the sentence. Words are output in their root form. Our (untested) assumption is that parsing will provide a better word analysis of the sentence than using a lightweight tagger. Parsing is performed using SAPIR, a broad-coverage chart parser (Harrison & Maxwell, 1986), that includes WordNet's lexicon within it. If a parse fails then partial parse fragments are collected, such that all the words in the sentence are accounted for in some way.

Second, words that are not in the four major part-of-speech categories (noun, verb, adjective, adverb) are dropped. Finally, any remaining stop words, pronouns, and a small number of "light verbs"[2] (verbs with little semantic weight) are dropped. For example:

> 916/120/574/1:
> **T:** The Christian Science Monitor newspaper on Monday pleaded for the release of American reporter Jill Carroll...
> **H:** Jill Carroll was abducted in Iraq.

will produce the bags:

> **T:** "Christian Science" "Monitor" "newspaper" "Monday" "plead" "release" "American" "reporter" "Jill" "Carroll" …
> **H:** "Jill" "Carroll" "abduct" "Iraq"

2.2 Knowledge-Based Lexical Comparison

BLUE-Lite's basic entailment mechanism is to see if each word in H is "entailed" by a word in T, e.g., "big" entails "large". This notion of "entailment" is

[2] These are: be, occur, happen, exist, start, end, begin, finish.

clearly weak as semantic arguments to (the predicate denoted by) the words are assumed to be the same, rather than checked through syntactic analysis. (We return to discuss this assumption later in Section 5). WordNet and DIRT are used extensively to determine word entailment as follows:

2.2.1 WordNet

WordNet (Fellbaum, 1998) is used to compare a T word and an H word in two stages:

1. **Collect possible word meanings (synsets):** For each word, BLUE-Lite collects possible synsets for it (i.e., the concepts it might denote).
2. **Search for an entailment relation:** The system then searches for an entailing WordNet relation between a synset of the T word and a synset of the H word.

In doing this, BLUE-Lite is considering all possible meanings of a word, and will conclude entailment if some meaning of the T word entails some meaning of the H word, on the grounds that if there is a relationship between senses of two words, then those senses were probably the senses intended by the author (a principle used by many word sense disambiguation algorithms (Navigli, 2009)).

1. Collect possible word meanings (synsets)

To identify the synsets corresponding to a word, we use several WordNet relations:

a. Basic Synset Collection: First, BLUE-Lite collects the synsets S that the word is a member of. The part of speech of the word is ignored for this, on the assumption that (usually) the same word with different parts of speech will have a similar meaning.

b. Morphosemantics: WordNet includes a morphosemantic database (Fellbaum, Osherson, and Clark, 2007) that relates noun senses and verb senses together distributed as a WordNet add-on ("standoff") file[3]. We use this to expand noun senses in the bag with their equivalent verb senses, and vice versa. The morphosemantic database includes entries such as:

build#v1 -agent→ maker#n1

stating that maker#n1 is the agent of build#v1 events. The identity relation - the one we are interested in here - is named "event", e.g.,:

build#v1 -event→ construction#n1

stating that construction#n1 denotes the actual event build#v1 (i.e. is a nominalization of the event itself). Thus, all noun and verb senses in the bag are augmented with their equivalent verb and noun senses, by following the "event" link in the morphosemantic database.

c. Pertains To: WordNet's "pertains-to" relation connects approximately equivalent (senses) of nouns, adjectives, and adverbs, e.g.,:

rapidly#r1 ←pertains-to→ quick#a1

All synsets pertained to by members of S are added to S.

d. Similar To: WordNet has a similar-to link relating adjective senses with approximately equivalent meaning, for example:

nice#a1 ←similar-to→ pleasant#s2

where "a" denotes adjective and "s" denotes a "similar adjective". ("s" is a somewhat redundant label stemming from historical development of WordNet). For any adjective sense in S, BLUE-Lite also adds the similar adjective senses to S using this relation. One special exception is made, namely to ignore cardinal#a2 (as in "cardinal number"), which is recorded as similar-to each integer in WordNet (one#s1, two#s1, etc.), although they are clearly not synonymous - this appears to be a semantic error in WordNet (v2.0).

2. Searching for an Entailment Relation

Given the synsets for a T word and an H word, as collected using the above method, BLUE-Lite then searches for an entailment relation from a T synset to an H synset. If one is found, then the T word is considered to entail the H word. Four WordNet entailment relations are used:

Equivalence: If the two words have a synset in common, they entail each other.

Hypernym: If an H sense is a hypernym (generalization) of a T sense, it is entailed, e.g.,

car#n1 −WN−isa→ vehicle#n1

[3] http://wordnet.princeton.edu/wordnet/download/standoff/

Part of: If a T sense is a part of an H sense, then the H sense is entailed. This is the mp (meronym part) relation in WordNet, e.g.,

Baghdad#n1 –WN-part-of→ Iraq#n1

Substance of: If a T sense is a substance of an H sense, then the H sense is entailed. This is the ms (meronym substance) relation in WordNet, e.g.,

snow#n2 –WN-substance-of→ snowball#n4

2.2.2 The DIRT Paraphrase Database

In addition to WordNet, BLUE-Lite also uses the DIRT database to search for an entailment relation between two words. DIRT itself (Lin and Pantel, 2001, Pantel et al., 2007) contains approximately 12 million paraphrases, discovered automatically from text, of the form:

(X *relation*$_1$ Y) ↔ (X *relation*$_2$ Y)

where relation is a dependency path between constitutents X and Y, and the implication denotes that the two patterns occurred in distributionally similar contexts. Empirically, this happens when the two patterns have similar meaning (i.e., are paraphrases), or when one implies the other (reflecting some general knowledge about the world). Two example DIRT rules are:

X loves Y ↔ X adores Y
X loves Y ↔ X has a passion for Y

In fact, a large number of the DIRT paraphrases are of the simple form:

X *verb* Y ↔ X *verb'* Y

Here, because the dependency paths in the condition and action are the same, 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-Lite uses these DIRT-derived inferential relationships when computing subsumption between words in the bags. Although this technique ignores paraphrases that use longer dependency paths, these longer path paraphrases seem, from informal experience, to be less reliable, and hence we believe we are exploiting the most reliable part of DIRT. Note that, unlike the WordNet comparisons, the DIRT comparisons are between words rather than between word senses.

An illustration of entailment using both WordNet and DIRT is:

916/120/574/1[4] (BLUE-Lite got this right):
T: ...Jill Carroll, seized in Baghdad...
H: Jill Carroll was abducted in Iraq.
via:
"Jill"(n) ↔ "Jill"(n)
"Carroll"(n) ↔ "Carroll"(n)
"abduct"(v) ←DIRT→ "seize"(v)
"Iraq"(n) ←WN-part-of- "Baghdad"(n)

2.2.3 Comparing Names

There is one important exception to simple word comparison in BLUE-Lite, namely the way it handles proper nouns (e.g., names). BLUE-Lite makes the assumption that two names (sequences of proper nouns) are coreferential if their last noun is the same. For example, "Barak Obama", "Obama", "President Obama", "Mr Barak Obama", etc. are all treated as coreferential as the last noun in the sequence is "Obama". Thus when comparing bags of words, if an H proper noun is not found in T, but *is* part of a proper noun sequence in the original H sentence whose last element *is* found in T, than that proper noun is deemed to match, on the grounds that it is part of a matching, multiword proper noun. An illustration is below:

909/214/72/10 (BLUE-Lite got this right):
T: Adams has made a direct appeal...
H: Gerry Adams made a direct appeal...
via:
"Adams" ↔ "Adams"
"Gerry" ↔ (extra part of the name "Adams")
"make" ↔ "make"
... ...

Here, the non-matching H word "Gerry" is recognized as part of the compound name "Gerry Adams", and hence treated as if it matched, rather than counted as a mismatch.

The policy of "same last name = same entity" can clearly go wrong, e.g., topic 912 in the development data had articles referring to both Cindy Shehan and Casey Shehan, sometimes causing corefe-

[4] This notation shows the IDs of Topic/H/Document/T

rence errors if both names occurred in the T-H pair. However, empirically, these problems appeared to be more the exception rather than the rule.

2.3 Context and Using Preceding Sentences

A characteristic of RTE6 is the need to use information from the textual context of T (i.e., the surrounding sentences) to help determine entailment. The most obvious case of this is for anaphora resolution, for example:

924/148/307/2 (BLUE-Lite got this right):
T-1: ...drug Vioxx...
T: Merck...pulled the...pain drug...
H: Vioxx is a pain drug.

Here, T in isolation does not entail H, but T in context does because the context shows that "the pain drug" in T refers to "Vioxx", mentioned in the preceding sentence T-1. In general, coreference resolution is very hard, in particular with newspaper articles where references can be oblique or even completely implicit, and often use very different wording to the referents they denote. Rather than attempting full coreference resolution, we adopted an approximate approach whereby if an H word was not entailed by any T word, BLUE-Lite looked (anywhere) in the preceding sentence for an entailing word with reduced confidence, on the grounds that an entity mentioned in the sentence preceding T might be (with reduced confidence) anaphorically or implicitly also part of the semantics of T itself.

2.4 Varying the Entailment Threshold

Finally we used a technique to vary the threshold used to conclude entailment on a per-topic basis. RTE6 includes 10 data sets, each about a quite different news topic. We found that, given a fixed entailment threshold, the number of entailments that BLUE-Lite found per topic varied quite considerably, although the actual number of entailments per topic is roughly constant (50-100), reflecting the fact that BLUE-Lite found some topics harder to find entailments in than others. This is illustrated in Figure 1, showing the F-score obtained at the following different thresholds for entailment:

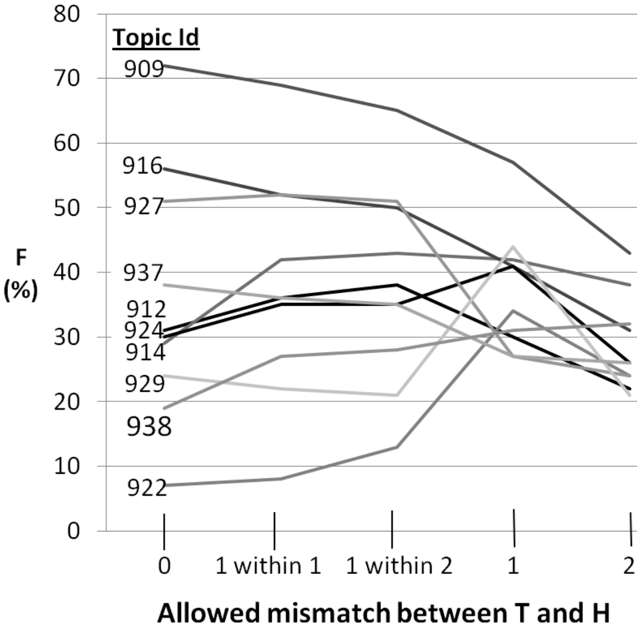

Figure 1: The optimal entailment threshold (allowed mismatches) varies considerably depending on the topic.

0 mismatches: all H words must be entailed by the T words.

1 mismatch within 1: all but one H word must be entailed by the T words, and the unmatched H word must be entailed by the sentence immediately preceding T.

1 mismatch within 2: all but one H word must be entailed by the T words, and the unmatched H word must be entailed by the two sentences immediately preceding T.

1 mismatch: all but one H word must be entailed by the T words.

2 mismatches: all but two H words must be entailed by the T words.

Note that each new threshold is progressively weaker than the previous, thus catching all the previous's entailments plus new ones.

The most important thing to note in Figure 1 is how much the curves vary: Some topics, e.g., topic 909, already have a high F-score at the strictest threshold, reflecting that even at that threshold a large number of entailments are found. Other topics, e.g., topic 922, have a poor F-score at the strictest threshold, reflecting that few entailments were found at that threshold, and that weakening it improves recall and hence F-score, reaching an optimum at a weaker threshold.

Although the number of entailments per topic found by BLUE-Lite varies (for a given threshold), the actual number of entailments in the training data per topic is approximately constant (about 50-100). Thus to help BLUE-Lite's performance in "difficult" topics, for each topic the threshold for concluding entailment was iteratively relaxed until a fixed minimum number N of entailments was found for that topic, with the aim of finding the threshold corresponding to the peak F-measure performance for that topic. The optimal value of N (=45) was found by generating a performance vs. N curve on the training data.

2.5 Search

Finally we note that finding the best pairing of T and H words is a search process, as each H word may be entailed by more than T word. As the H sentences are relatively short, BLUE-Lite does an exhaustive search to find the best pairings.

3. Evaluation

We now present the results of our experiments, followed by a more detailed failure analysis. In the experiments, we compared performance using three fixed degrees of "strictness" for concluding entailment as follows:

- **0 mismatches:** all H words must be entailed by the T words.
- **1 mismatch within 1:** all but one H word must be entailed by the T words, and the unmatched H word must be entailed by the sentence immediately preceding T.
- **1 mismatch:** all but one H word must be entailed by the T words.

and also the variable entailment threshold, as described earlier in Section 2.4:

- **Variable:** Chose the most restrictive threshold per topic (0 mismatches, 1 mismatch within 1, or 1 mismatch) that also produces at least N (45) entailments.

We also ran the system ablating DIRT, WordNet (WN), and both, to measure the effects of each knowledge resource. The overall results (microaveraged F-measure, shown as a percentage) are shown in Table 1.

<u>Performance on RTE6 Development Set</u>

Knowledge Sources ↓	**Entailment Threshold** (number of mismatches allowed)			
	0	1 within 1	1	Variable
none	21.81	26.24	39.72	39.72
DIRT	25.07	31.06	39.72	40.45
WN	31.81	36.36	39.69	43.27
WN+DIRT	37.59	40.30	35.29	42.68

<u>Performance on RTE6 Test Set</u>

Knowledge Sources ↓	**Entailment Threshold** (number of mismatches allowed)			
	0	1 within 1	1	Variable
none	23.35	27.66	40.35	40.56
DIRT	25.47	30.72	40.55	39.57
WN	35.44	39.41	38.68	40.02
WN+DIRT	37.20	41.56	38.74	43.99

Table 1: BLUE-Lite's performance on RTE6 (Micro-averaged F-measure, shown as a percentage)

There are several interesting items of note. First, a striking feature of these tables is the third column: The simple strategy of a bag-of-words match, with 1 mismatch is allowed, produces an F-measure of around 40%, even without WordNet and DIRT. Although there is some selectivity used in picking the "bag of words" (Section 2.1), and names are compared in a special way, it is still surprising this strategy does so (relatively) well, even when words are compared using simple equality (no WordNet or DIRT).

Second, the data shows that BLUE-Lite is able to improve on this F-meaure by using DIRT, WordNet, and either a "1 mismatch within 1" or "variable" entailment threshold. Although the trends are not completely uniform, the "1 mismatch within 1" threshold scored around 41%, while the the variable (topic-specific) threshold scored around 43%-44%. The optimal configuration of the system was using WordNet, a variable entailment theshold, and (on the test data) DIRT all together.

Third, varying the entailment threshold (the last column in Table 2) generally produced the best results. The particular way the variability is controlled is a little unsatisfying as it exploits an idiosyncracy of the RTE6 data, namely that the total number of entailments per topic is approximately constant (around 50-100). However, even with real world problems, one might still be able to estimate how hard it is to find an entailment (e.g., shorter sentences may be easier to entail than longer ones; common words may be easier to entail than rare words), and hence be able to vary the entailment threshold based on that guess. Thus at a general level, varying the entailment threshold based on characteristics of the problem and/or domain seems to be a technique that is potentially portable to other entailment settings, and worth exploring further.

A useful way of understanding these results, in particular the low results using a stricter threshold for entailment, is that the first task for BLUE-Lite is to simply find enough entailments. Requiring all H words to match T (the first column, 0 mismatch) is too strict a criterion, producing too low a recall in most topics and hence low F-measures. However, when this criterion is relaxed so that enough entailments are found, the focus then becomes one of ordering those entailments well. The results suggest that WordNet and DIRT together provide some benefits for this ordering, and that varying the entailment threshold provides a better means for choosing a cutoff point along that ordering.

4. Failure Analysis

To understand BLUE-Lite's performance more, we examined cases where BLUE-Lite made a mistake, in particular boundary cases where BLUE-Lite strongly (perfect match) concluded entailment for a non-entailing pair, and where BLUE-Lite strongly (large mismatch) rejected entailment for an entailing pair. We identified twelve broad classes of problems, as follows:

4.1 Order-Independent Analysis of T and H

Clearly ignoring syntactic structure inherently limits BLUE-Lite's performance; just because H has the same (or entailed) words in it as T does not mean those words will be related in similar ways, or even at all. Two examples of such failures are:

912/74/569/10 (BLUE-Lite wrong, saying YES):
T: ...the mother of a ... Marine killed in Iraq...sided...with Sheehan.
H* [5]**:** …Sheehan was killed in Iraq.

937/259/608/17 (BLUE-Lite wrong saying YES):
T: Whittington owns property in...Austin...
H*: …Whittington is from Austin.

In both these cases, H is clearly not entailed by T, even though it contains the same words. These kinds of errors are more common if one mismatch is allowed, e.g.,:

929/11/614/3 (BLUE-Lite wrong, saying YES):
T: Texas is also looking for states to house 360 refugees...Passey told reporters...
H*: A state of emergency was declared…
"state"(n) ↔ "state"(n)
"emergency"(n) ↔ [no match]
"declare"(v) ←WN-isa- "tell"(v)

In this case, the two occurences of the word "state" have been treated as equivalent, although here their senses are completely diffferent.

However, although these examples clearly illustrate the inherent limitations of ignoring word order and syntax, in practice these problems are less common than one might expect. We discuss this in more detail later in Section 5.

4.2 Use of WordNet

Table 1 illustrates that WordNet provides some leverage for us, contributing an additional 0%-10% in F measure depending on the system configuation. A typical example of WordNet's use is:

924/153/497/10 (Blue got this right):
T: Merck pulled ...Vioxx off the market...
H: Merck withdrew Vioxx from the market.
via:
"withdraw"(v) ←WN-isa- "pull"(v)

WordNet-related errors occurred primarily when BLUE-Lite found a WordNet connection between words that were clearly unrelated in the sentences themselves. Although this is primarily a problem with BLUE-Lite's lexical approach rather than WordNet itself, WordNet can exacerbate the problem by enabling more matches to be found, both good and bad. For example, WordNet found an incorrect match between "destruction" and "kill" in the example below:

[5] We use the notation "**H***" to denote an H that is NOT entailed by T.

912/74/2/9 (BLUE-Lite wrong, saying YES):
T: ...Iraq...had weapons of mass destruction...Shehan...
H*: Casey Sheehan was killed in Iraq.
via:
"kill"(v) ←WN-synonym→ "destruction"(adj)

In addition, occasionally some obscure or questionable word senses in the WordNet database (version 2.0) caused errors, for example:

924/148/309/17 (BLUE-Lite wrong saying YES):
T: … Vioxx could have an impact on...the drug....
H*: Vioxx is a pain drug.
via:
"pain"(n) ←WN-isa- "have"(v)

where the WordNet sense of "have" used was:

have#v12: "have": suffer from; be ill with; as in "She has arthritis".

This is clearly a dubious sense to assign to "have".

Overall, the primary advantage of WordNet was its ability to frequently connect related words that were lexically dissimilar but semantically related via entailment. Some additional, interesting, good examples of this from the development set runs were:

"market"(n) ← "sale"(n)
"study"(n) ↔ "report"(v)
"increase"(v) ← "elevate"(v)
"fall"(v) ↔ "decline"(v)
"misleading"(adj) ↔ "false"(adj)
"US"(n) ↔ "American"(adj)
"stock"(adj) ← "share"(n)

Although this mechanism can clearly go wrong often, it seemed to produce an overall net benefit as illustrated earlier in Table 1.

4.3 Use of DIRT

DIRT, like WordNet, results in additional entailment relations between words. A good example of DIRT's use is:

916/120/574/1 (BLUE-Lite got this right):
T: ...Jill Carroll, seized in Baghdad...
H: Jill Carroll was abducted in Iraq.
via:
abduct"(v) ←DIRT→ "seize"(v)

Although many of the DIRT-derived relationships duplicate WordNet's, there are novel ones such as the above that can aid entailment. However, there are also numerous bad relationships, sometimes leading to bad entailments, for example:

924/153/497/9 (BLUE-Lite wrong, saying YES):
T: Merck is redeploying...marketing...dedicated to Vioxx...
H*: Merck withdrew Vioxx from the market.
via:
"withdraw"(v) ←DIRT→ "redeploy"(v)

Some additional good examples of DIRT equivalences used on the development data are:

"run"(v) ↔ "oversee"(v)
"mark"(v) ↔ "symbolize"(v)
"say"(v) ↔ "report"(n)
"shoot"(v) ↔ "injure"(v)

while some bad examples are:

"remember"(v) ↔ "expect"(v)
"deliver"(v) ↔ "make"(n)
"withdraw"(v) ↔ "back"(v)
"shoot"(v) ↔ "get"(v)

Our results in Table 1 are somewhat inconclusive about the overall effect of DIRT; in some cases it slightly improves performance, while in others it slightly harms it. Further work is needed to better exploit this resource.

4.4 Missing Paraphrases

Several failures were due to T and H expressing the same knowledge in significantly different ways, i.e., beyond simple paraphrase equivalences that DIRT or WordNet were able to recognize. For example, recognizing entailment in:

909/214/397/5 (BLUE-Lite wrong, saying NO):
T: ...Gerry Adams...appealed to IRA members to leave behind their "armed struggle" in favor of democratic politics.
H: Gerry Adams made a direct appeal to the Irish Republican Army to embrace purely democratic activity.

requires recognizing the equivalence of both

"appealed" ↔ "made a direct appeal"
"in favor of democratic politics"
↔ "embrace purely democratic activity"

The first equivalence might be generalized and encoded as a rule that "*verb*" ↔ "make a *verb-nominalization*". The second requires knowledge that favoring something implies (metaphorically) embracing it. Knowing additional phrasal equivalences like this would help in determining entailment, although they would also necessitate structural analysis of the sentences to apply them.

4.5 Commonsense Knowledge

Closely related to paraphrasing is the need for commonsense knowledge to conclude entailment. An interesting example is:

> 922/174/262/12 (BLUE-Lite wrong, saying NO):
> **T:** … the provision allowing secret searches of homes, businesses and personal property.
> **H:** … provisions let police conduct secret searches of people's homes or businesses.

Concluding entailment here includes realizing that, among other things:

> "homes" (T) -entails→ "people's homes" (H)
> "allow searches" (T)
> -entails→ "let police conduct searches" (H)

The first of these requires using the general knowledge that people live in homes. The second requires knowledge that "allow X" entails "letting Y conduct/do X", and that in this context (statutes about law enforcement) it is the police that are doing the searching.

Three other interesting examples are:

> 929/5/664/3 (BLUE-Lite wrong, saying NO):
> **T:** Local schools have already closed...amid fears the hurricane could strike...
> **H:** Texas braced for Hurricane Rita.

requiring knowledge that closing schools is a common means of bracing for a hurricane, and:

> 927/229/6/29 (BLUE-Lite wrong, saying NO):
> **T:** Jennings anchored ABC's evening news for two years...
> **H:** Peter Jennings delivered the news to Americans each night.

requiring knowledge that anchoring the news involves delivering the news every night. (This example also requires knowing that ABC is American, and that an American broadcasting company will broadcast to Americans). Finally:

> 938/303/346/10 (BLUE-Lite wrong, saying NO):
> **T:** …the foundation will be a fund-raising organization.
> **H:** The World Trade Center Memorial Foundation was created to raise money.

requires knowing that fund-raising implies raising money, and that (in this context) "will be" implies an intensional act ("was created to").

4.6 Geography & Other Factual Knowledge

Some pairs required geographical knowledge, for example:

> 914/61/771/3 (BLUE-Lite wrong, saying NO):
> **T:** ....explosions rocked tourist camps at Shitani and Ras Soltan....south of Taba.
> **H:** ...blasts targeted tourist resorts in Sinai.

This example requires knowing that Shitani, Ras Soltan, or Taba are in the Sinai. WordNet does contain some limited geographical knowledge (e.g., that Baghdad is part of Iraq), but not at this level of detail. More systematic use of factual world knowledge would be very helpful here, or other parts of the T text could be used to identify the required information.

4.7 Arithmetic

A few pairs required simple arithmetic, outside the scope of BLUE-Lite and a good example of a phenomenon clearly requiring more than a lexical solution. For example:

> 914/61/771/3 (BLUE-Lite wrong, saying NO):
> **T:** Shortly after the hotel blast, another two explosions rocked tourist camps...
> **H:** Three blasts targeted tourist resorts...

Here "the blast" (T) + "two explosions" (T) = "Three blasts" (H), a reasonably complex task for an entailment system to recognize and perform.

4.8 Dates and Calendrics

Date references can be difficult to align, and clearly a bag-of-words approach will be inadequate when the date is split into disconnected fragments. Similarly, relative dates (e.g., "last Tuesday") are complex and outside the capabilities of a bag-of-words approach. A particularly difficult example exhibiting both is:

> 924/154/380/2 (BLUE-Lite wrong, saying NO):
> **T:** Merck announced a global withdrawal Thursday of Vioxx...
> **H:** Merck pulled Vioxx off the market on September 30, 2004.

This example requires realizing that "Thursday" (T) refers to the Thursday before the article date (Oct 1st 2004), which (from a calendar) happens to be September 30th, 2004. It also requires recognizing "September 30, 2004" as a date, rather than breaking it up (as BLUE-Lite does) into the bag {"September" "30" "2004"}.

4.9 Modals and Negation

Modals and negation can change the polarity of entailment (Nairn et al., 2006). However these phenomena cannot be handled without structural analysis, and hence they caused occasional errors for BLUE-Lite, for example:

> 924/153/309/23 (BLUE-Lite wrong saying YES):
> **T:** Merck announced that it would withdraw Vioxx from the market...
> **H*:** Merck withdrew Vioxx from the market.

Here, announcing X does not imply X itself, something that lexical matching is unable to detect and handle. (However, these phenomena seemed rare in the RTE6 data).

4.10 Cataphora (Forward Reference)

To handle anaphoric (backward) reference (in a limited way), BLUE-Lite looks at the sentence preceding T for a match if an H word does not match T. However, there are a few examples of cataphora (forward reference) where the referent occurs *after* T, e.g.,

> 912/74/2/1 (BLUE-Lite wrong, saying NO):
> **T:** The angry mother of a fallen U.S. soldier...
> **T+2:** Cindy Sheehan told reporters...
> **T+6:** Her son, Casey, 24, was killed in…Iraq...
> **H:** Casey Sheehan was killed in Iraq.

In this case, the reference "The angry mother" is resolved two sentences later (T+2), and the reference "a fallen U.S. soldier" is resolved 6 sentences later (T+6). This is well outside the scope of BLUE-Lite.

4.11 Morphology

Occasionally, morphological variation caused entailments to be missed, for example:

> 937/258/644/1 (BLUE-Lite wrong, saying NO):
> **T:** a 78-year-old hunting partner...
> **H:** Harry M. Whittington is 78 years old.

In this case, BLUE-Lite treated "78-year-old" as a single lexical item, that then did not entail the three items {"78" "years" "old"} in H.

4.12 Idioms

Occasionally, idioms result in very different phrasing of entailing sentences, e.g.,:

> 938/320/211/6 (BLUE-Lite wrong, saying NO):
> **T:** The pools will be centered within, but slightly smaller than, the tower outlines.
> **H:** The pools mark the footprints of the Twin Towers.

BLUE-Lite was unable to handle this example.

5 Coherence and Syntactic Structure

5.1 The Coherence Conjecture

As discussed earlier, it is somewhat surprising that "knowledge-based lexical entailment" works at all, given the volume of apparently important syntactic information that it ignores. We had a similar result last year where BLUE's use of syntax resulted in only a small (1%) improvement over using its knowledge-based bag-of-words module alone. Why is this, and why is it so hard to gain substantial advantage from syntactic/structural analyses of the sentences? A conjecture is as follows:

> **The Coherence Conjecture**
>
> If T and H are coherent (plausible, not nonsensical, consistent with commonsense knowledge, topically similar), and if two words in H also appear in T, then it is very likely that the semantic relationship between those two words in H, and the semantic relationship between those words in T, are the same.

The justification for this conjecture is that, if T and H are coherent, then this massively constrains the possible relationships that can exist between the two words, often so much that there is only one key relationship, and hence a system can reliably assume the relationship in T and in H are the same, without explicit structural/syntactic analysis to confirm it. To the extent that the Coherence Conjecture is true, we only need to find the H words in T; if the words are there, then it is a priori likely that the semantic relationship between the words is there also there, even without checking this through syntactic analysis.

Clearly, changing the word order changes the semantic relationship. But the important point here is that, in general, changing the the word order produces incoherent sentences, thus violating the premise of the Coherence Conjecture. For example, clearly T does not entail H in:

T: Einstein discovered relativity.
H*: Relativity discovered Einstein.

But the important thing here is that an H like this will rarely occur in the kinds of "real life" settings that RTE has been considering to date, as H is incoherent.

The constraint on possible relationships is even stronger in RTE6, as both T and H are not only coherent, but also true in the world (as they are drawn from newspaper articles, to the extent that newspapers report the truth). Thus we do not see Hs in RTE6 that are false in the world such as:

H*: Cindy Sheehan was killed in Iraq.
[It was her son, Casey]
H*: Gerry Adams is the prime minister of Ireland.
[It is (was) Bertie Ahern]
H*: The European Medicines Agency is the maker of Vioxx. [It is Merck]

Such Hs would cause havoc for lexical approaches such as ours. However, as they are false in the world, they are not stated in newspaper articles, and therefore do not appear as hypotheses in the summarization setting of RTE6.

5.2 Coherence and Search Tasks

In a more general search setting for RTE, however, where users are asking queries, the degree of coherence in the query may be less. For "true/false" questions (which can be answered using RTE technology by turning them into assertions) the user's query may in fact be false in the world, e.g.,:

H: Cindy Sheehan was killed in Iraq.

Although such queries increase the need for syntactic analysis, the Coherence Conjecture still applies; for example, we would not expect a question such as:

H: Iraq was killed in Cindy Sheehan.

as such a query would be incoherent. Again, coherence strongly constrains the allowable semantic relations between constituents.

A more challenging search setting for RTE is to answer "find a value" questions. Such questions can also be answered using RTE technology by substituting a linguistic variable (e.g., "someone", "something") for the queried item, and then turning the question into an assertion (H), for example:

Who is the prime minister of Ireland?

can be transformed into:

H: Someone is the prime minister of Ireland.

An RTE system's task is then to not only return an entailment decision, but also (for positive entailments) return the item that the linguistic variable (here, "someone") matched in the entailing text -- in this case, hopefully the name of the Irish Prime Minister. Clearly in this setting, structural analysis is essential; a bag of words approach would consider any word in a sentence mentioning "prime minister of Ireland" to be a solution, clearly not a desirable outcome.

6. Conclusion

Our RTE6 system, BLUE-Lite, can be characterized as performing "knowledge-based lexical entailment", using DIRT and WordNet as knowledge sources, using the sentence preceding T as context, and varying the entailment threshold depending on the topic. Our experiments show that a combination of all these features performs relatively well (F=44%), and also that a fairly simple baseline that uses none of these techniques, namely that all but one of the H words are in T, performs moderately well also (F=40%). Our failure analysis shows that a large number of problem types remain, most of which require reintroducing structural/syntactic analysis of the sentences. We have also conjectured that the lexical approach does relatively well because the sentences are coherent, thus heavily constraining the possible semantic relationships that will exist between words, limiting the additional value of explicit relational (structural) analysis.

Ultimately, we would like to build RTE systems with much higher performance than those of today. While BLUE-Lite's lexical approach performed well in relative terms, it still performed poorly in absolute terms -- we would like to do considerably better than an F-measure of 44%. Given the inherent upper bound on purely lexical methods, it is essential to ultimately reintroduce structural information, in particular to address some of the major challenges described earlier of paraphrasing, reasoning, and commonsense knowledge. There is still a wealth of opportunities available for improvement and further progress, and much work to be done.

References:

Clark, P., Harrison, P. An Inference-Based Approach to Recognizing Entailment. In Proc TAC'09. 2009.

Condoravdi, C.,Karttunen, L. Computing Relative Polarity for Textual Inference. In Proc ICoS-5 (Inference in Computational Semantics). 2006.

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

Fellbaum, C., Osherson, A., Clark, P. Putting Semantics into WordNet's "Morphosemantic" Links In: *Proc 3rd Language & Technology Conference (LTC'07)*, Poznan, Poland, 2007.

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

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

Navigli, R. Word Sense Disambiguation: A Survey. In: ACM Computing Surveys, 41 (2), Feb 2009.

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