# The Role of Sentence Structure in Recognizing Textual Entailment

**Catherine Blake**
School of Information and Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3360
`cablake@email.unc.edu`

## Abstract

Recent research suggests that sentence structure can improve the accuracy of recognizing textual entailments and paraphrasing. Although background knowledge such as gazetteers, WordNet and custom built knowledge bases are also likely to improve performance, our goal in this paper is to characterize the syntactic features alone that aid in accurate entailment prediction. We describe candidate features, the role of machine learning, and two final decision rules. These rules resulted in an accuracy of 60.50 and 65.87% and average precision of 58.97 and 60.96% in $RTE3_{Test}$ and suggest that sentence structure alone can improve entailment accuracy by 9.25 to 14.62% over the baseline majority class.

## 1 Introduction

Understanding written language is a non-trivial task. It takes years for children to read, and ambiguities of written communication remain long after we learn the basics. Despite these apparent complexities, the bag-of-words (BOW) approach, which ignores structure both within a sentence and within a document, continues to dominate information retrieval, and to some extent document summarization and paraphrasing and entailment systems.

The rational behind the BOW approach is in part simplicity (it is much easier and less computationally expensive to compare terms in one sentence with terms in another, than to generate the sentence structure); and in part accuracy, the BOW approach continues to achieve similar if not improved performance than information retrieval systems employing deep language or logical based representations. This performance is surprising when you consider that a BOW approach could not distinguish between the very different meaning conveyed by: `(1)Slow down so that you don't hit the riders on the road` and `(2)Don't slow down so you hit the riders on the road.` A system that employed a syntactic representation of these sentences however, could detect that the `don't` modifier applies to `hit` in first sentence and to `slow` second.

In contrast to information retrieval, researchers in paraphrase and entailment detection have increased their use of sentence structure. Fewer than half of the submissions in the first Recognizing Textual Entailment challenge (RTE1) employed syntax (13/28, 46%) (Dagan, Glickman, & Magnini, 2005), but more than two-thirds (28/41, 68%) of the second RTE challenge (RTE2) submissions employed syntax (Bar-Haim et al., 2006). Furthermore, for the first time, the RTE2 results showed that systems employing deep language features, such as syntactic or logical representations of text, could outperform the purely semantic overlap approach typified by BOW. Earlier findings such as (Vanderwende, Coughlin, & Dolan, 2005) also suggest that sentence structure plays an important role in recognizing textual entailment and paraphrasing accurately.

Our goal in this paper is to explore the degree to which sentence structure alone influences the accuracy of entailment and paraphrase detection.



*Proceedings of the Workshop on Textual Entailment and Paraphrasing*, pages 101–106,
Prague, June 2007. 

Other than a lexicon (which is used to identify the base form of a term), our approach uses no background knowledge, such as WordNet (Miller, 1995), extensive dictionaries (Litkowski, 2006) or custom-built knowledge-bases (Hickl et al., 2006) that have been successfully employed by other systems. While such semantic knowledge should improve entailment performance, we deliberately avoid these sources to isolate the impact of sentence structure alone.

## 2 System Architecture

### 2.1 Lexical Processing

Our approach requires an explicit representation of structure in both the hypothesis (HSent) and test (TSent) sentence(s). Systems in RTE challenges employ a variety of parsers. In RTE2 the most popular sentence structure was generated by Minipar (Lin, 1998), perhaps because it is also one of the fastest parsers. Our system uses the typed dependency tree generated by the Stanford Parser (Klein & Manning, 2002). A complete set of parser tags and the method used to map from a constituent to a typed dependency grammar can be found in (de Marneffe et al., 2006). Figure 1 shows an example typed dependency grammar for pair id 355 in the $\text{RTE3}_{\text{Test}}$ set.

### 2.2 Lexicon

Our proposed approach requires the base form of each term. We considered two lexicons for this purpose: WordNet (Miller, 1995) and the SPECIALIST lexicon (National Library of Medicine, 2000). The latter is part of the National Library of Medicine's (NLM) Unified Medical Language System (UMLS) and comprises terms drawn from medical abstracts, and dictionaries, both medical and contemporary.

With 412,149 entries, the SPECIALIST lexicon (version 2006AA) is substantially larger than the 5,947 entries in WordNet (Version 3.0). To understand the level of overlap between the lexicons we loaded both into an oracle database. Our subsequent analysis revealed that of the WordNet entries, 5008 (84.1%) had a morphological base form in the SPECIALIST lexicon. Of the 548 distinct entries that differed between the two lexicons, 389 differed because either the UMLS (214 terms) or WordNet (11 terms) did not have a base form. These results suggest that although the NLM did not develop their lexicon for news articles, the entries in the SPECIALIST lexicon subsumes most terms found in the more frequently used WordNet lexicon. Thus, our system uses the base form of terms from the SPECIALIST lexicon.

### 2.3 Collapsing Preposition Paths

Previous work (Lin & Pantel, 2001) suggests the utility of collapsing paths through prepositions. The type dependency does have a preposition tag, prep, however, we found that the parser typically assigns a more general tag, such as dep (see the dep tag in Figure 1 between wrapped and by). Instead of using the prep tag, the system collapses paths that contain a preposition from the SPECIALIST lexicon. For example, the system

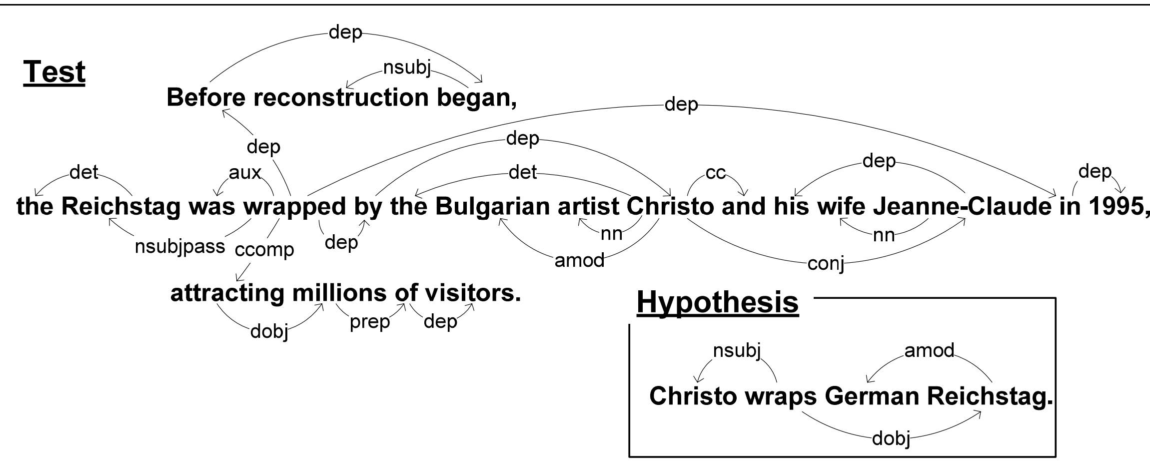


**Figure 1.** Dependency grammar tree for pair identifier 355 in the $\text{RTE3}_{\text{Test}}$

collapses four paths in $\text{TSent}_{\text{EG}}$ millions of visitors, wrapped in 1995, wrapped by Christo, and wrapped began before.

### 2.4 Base Level Sentence Features

The typed dependency grammar, such as that shown in Figure 1, can produce many different features that may indicate entailment. Our current implementation uses the following four base level features.

(1) Subject: The system identifies the subject(s) of a sentence using heuristics and the parser subject tags nsubjpass and nsubj.
(2) Object: The system uses the parser tag dobj to identify the object(s) in each sentence.
(3) Verb: The system tags all terms linked with either the subject or the object as a verb. For example, wrapped is tagged as the verb wrap from the link wrapped nsubjpass Reichstag shown in Figure 1.
(4) Preposition: As described in section 2.3 the system collapses paths that include a preposition.

The subject feature had the most coverage of the base level features and the system identified at least one subject for 789 of the 800 hypotheses sentences in $\text{RTE3}_{\text{Devmt}}$. We wrote heuristics that use the parser tags to identify the subject of the remaining 11 sentences. The system found subjects for seven of those eight remaining hypothesis sentences (3 were duplicate sentences). In contrast, the object feature had the least coverage, with the system identifying objects for only 480 of the 800 hypotheses in the RTE3 revised development set ($\text{RTE3}_{\text{Devmt}}$).

In addition to the head noun of a subject, modifying nouns can also be important to recognize entailment. Consider the underlined section of $\text{TSent}_{\text{EG}}$: which was later bought by <u>the Russian state-owned oil company Rosneft</u>. This sentence would lend support to hypotheses sentences that start with The Baikalfinasgroup was bought by … and end with any of the following phrases an oil company, a company, Rosneft, the Rosneft Company, the Rosneft oil company, a Russian company, a Russian Oil company, a state-owned company etc. Our system ensures the detection of these valid entailments by adding noun compounds and all modifiers associated with the subject and object term.

### 2.5 Derived Sentence Features

We reviewed previous RTE challenges and a subset of $\text{RTE3}_{\text{Devmt}}$ sentences before arriving at the following derived features that build on the base level features described in 2.4. The features that use 'opposite' approximate the difference between passive and active tense. For each hypothesis sentence, the system records both the number of matches (#match), and the percentage of matches (%match) that are supported by the test sentence(s).

(1) Triple: The system compares the subject-verb-objects in HSent with the corresponding triple in TSent.
(2) Triple Opposite: The system matches the verbs in both HSent and TSent, but matches the subject in HSent with the object in TSent.
(3) Triple Subject Object: This feature approximates the triple in (1) by comparing only the subject and the object in HSent with TSent, but ignoring the verb.
(4) Triple Subject Object Opposite: The system compares the objects in HSent with the subjects in TSent.
(5) Subject Subject: In addition to the triples used in the derived features 1-4, the system stores subject-verb and object-verb pairs. This feature compares the distinct number of subjects in HSent with those in TSent.
(6) Verb Verb: The system compares only the verb in the subject-verb, object-verb tuples in HSent with those in TSent.
(7) Subject Verb: The system compares the distinct subjects in HSent with the distinct verbs in TSent.
(8) Verb Subject: The system compares the verb in HSent with the subject in TSent.
(9) Verb Preposition: The system compares both the preposition and verb in HSent with those in TSent.
(10) Subject Preposition: The system compares both the subject and preposition in HSent with those in TSent.
(11) Subject Word: The system compares the distinct subjects in HSent with the distinct words in TSent. This is the most general of all 11 derived features used in the current system

## 2.6 Combining Features

A final decision rule requires a combination of the derived features in section 2.5. We used both previous RTE challenges and machine learning over the derived features to inform the final decision rules. For the latter, we chose a decision tree classifier because in addition to classification accuracy, we are also interested in gaining insight into the underlying syntactic features that produce the highest predictive accuracy.

The decision trees shown in Figure 2 were generated using the Oracle Data Miner 10.2.0.2. Tree (A) suggests that if there is less than a 63.33% similarity between the number of subjects in the hypothesis sentence and the words in any of the test sentences (feature 11), that the hypothesis sentence is not entailed by the test sentence(s). The NO prediction from this rule would be correct in 71% cases, and assigning NO would apply to 42% of sentences in the development set. A YES prediction would be correct in 69% of sentences, and a YES prediction would take place in 57% of sentences in the development set. Tree (B) also suggests that an increase in the number of matches between the subject in the hypothesis sentence and the words used in the test sentence(s) is indicative of an entailment.

```
Decision Tree (A)
  If Subject-Word match <= 63.3%

        NO                YES
  Accuracy=71%      Accuracy=69%
  Coverage=43%      Coverage= 57%

Decision Tree (B)
   If Subject-Word match <= 1.5

        NO                YES
   Accuracy=58%     Accuracy=64%
   Coverage=52%     Coverage=48%
```

**Figure 2.** Decision trees generated for the revised $RTE3_{Devmt}$ set during decision rule development.

Although tempting to implement the decision tree with the highest accuracy, we should first consider the greedy search employed by this algorithm. At each level of recursion, a decision tree algorithm selects the single feature that best improves performance (in this case, the purity of the resulting leaves, i.e. so that sentences in each leaf have all YES or all NO responses).

Now consider feature 1, where the subject, verb and object triple in the hypothesis sentence matches the corresponding triple in a test sentence. Even though the predictive accuracy of this feature is high (74.36%), it is unlikely that this feature will provide the best purity because only a small number of sentences (39 in $RTE3_{Devmt}$) match. Similarly, a subject-object match has the highest predictive accuracy of any feature in $RTE3_{Devmt}$ (78.79%), but again few sentences (66 in $RTE3_{Devm}$) match.

## 2.7 Final Decision Rules

We submitted two different decision rules to RTE3 based on thresholds set to optimize performance in $RTE3_{Devmt}$ set. The thresholds do not consider the source of a sentence, i.e. from information extraction, summarization, information retrieval or question answering activities.

The first decision rule adds the proportion of matches for each of the derived features described in section 2.5 and assigns YES when the total proportion is greater than or equal to a threshold 2.4. Thus, the first decision rule overly favors sentences where the subject, verb and object match both HSent and TSent because if a sentence pair matches on feature 1, then the system also counts a match for features 3, 4, 5, and 8. This lack of feature independence is intentional, and consistent with our intuition that feature 1 is a good indicator of entailment.

To arrive at the second decision rule, we considered the features proposed by decision trees with a non-greedy search strategy that favors high quality features even when only a small percentage of sentences match. The second rule predicts YES under the following conditions: when the subject, verb, and object of HSent match those in any TSent (feature 1), in either order (feature 2) or when the subject and object from the HSent triple match any TSent (feature 3), or when the TSent subject matches >= 80% of the HSent subject terms (feature 5) or when the TSent subject and preposition matches >=70% of those in HSent (feature 10) or when TSent word matches >= 70% of the subject terms in the HSent sentence (feature 11).

| Feature | | RTE3$_{Devmt}$ | | | RTE3$_{Test}$ | | | RTE2$_{All}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Total | Pos | %Accy | Total | Pos | %Accy | Total | Pos | %Accy |
| 1 | Triple | 35 | 26 | 74.29 | 37 | 24 | 64.86 | 47 | 35 | 74.47 |
| 2 | Triple Opposite | 4 | 3 | 75.00 | 9 | 4 | 44.44 | 2 | 1 | 50.00 |
| 3 | Triple Subj Obj | 66 | 52 | 78.79 | 76 | 47 | 61.84 | 102 | 69 | 67.65 |
| 4 | Triple Subj Obj Opp. | 9 | 4 | 44.44 | 16 | 7 | 43.75 | 10 | 5 | 50.00 |
| 5 | Subject-Subject | 750 | 397 | 52.93 | 760 | 404 | 53.16 | 777 | 391 | 50.32 |
| 6 | Verb-Verb | 330 | 196 | 59.39 | 345 | 181 | 52.46 | 395 | 208 | 52.66 |
| 7 | Subject-Verb | 297 | 178 | 59.93 | 291 | 168 | 57.73 | 292 | 154 | 52.74 |
| 8 | Verb-Subject | 348 | 196 | 56.32 | 369 | 207 | 56.10 | 398 | 212 | 53.27 |
| 9 | Verb-Preposition | 303 | 178 | 58.75 | 312 | 167 | 53.53 | 355 | 190 | 53.52 |
| 10 | Subject-Preposition | 522 | 306 | 58.62 | 540 | 310 | 57.41 | 585 | 303 | 51.79 |
| 11 | Subject-Word | 771 | 406 | 52.66 | 769 | 407 | 52.93 | 790 | 395 | 50.00 |

**Table 1.** Coverage and accuracy of each derived feature for RTE3 revised development collection (RTE3$_{Devmt}$), the RTE3 Test collection (RTE3$_{Test}$ ) and the entire RTE2 collection (RTE2$_{All}$).

## 3 Results

The experiments were completed using the revised RTE3 development set (RTE3$_{Devmt}$) before the RTE3$_{Test}$ results were released. The remaining RTE2 and RTE3$_{Test}$ analyses were then conducted.

### 3.1 Accuracy of Derived Features

Table 1 shows the accuracy of *any* match between the derived features described in section 2.5. Complete matching triples (feature 1), and matching subjects and objects in the triple (feature 2) provide the highest individual accuracy.

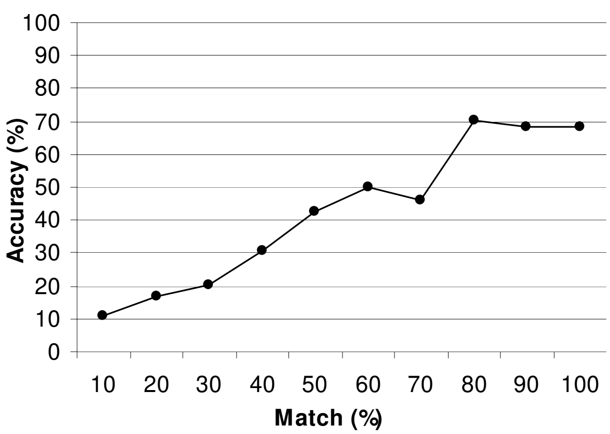


**Figure 3.** Correlation between accuracy and the percentage subjects in HSent that have a corresponding subject in TSent (feature 5).

The results in Table 1 do not consider the degree of feature match. For example, only one of the words from TSent in sentence 525's (RTE3$_{Devmt}$) matched the eight subject terms in corresponding HSent. If the derived features outlined in section 2.7 did capture the underlying structure of an entailment, you would expect an increased match would correlate with increased accuracy. We explored the correlations for each of the derived features. Figure 3 suggests entailment accuracy increases with an increase in the percentage of TSent subject terms that match HSent terms. (feature 5) and demonstrates why we set the 80% threshold for feature 5 in the second decision rule.

### 3.2 Accuracy of Decision Rules

Of the 800 sentences in RTE3$_{Devmt}$, the annotators labeled 412 as an entailment. Thus, without any information about HSent or TSent, the system would assign YES (the majority class) to each sentence, which would result in 51.50% accuracy.

The first decision rule considers the total percentage match of all features defined in section 2.5. We arrived at a threshold of 2.4 by ranking the development set in decreasing order the total percentage match and identifying where the threshold would lead to an accuracy of around 65%. Many sentences had a threshold of around 2.4, and the overall accuracy of the first decision on the RTE3$_{Devmt}$ set was 62.38%, compared to 60.50% in RTE3$_{Test}$. We consider the first decision rule a baseline and the second rule is our real submission.

The second rule uses only a sub-set of the derived features (1, 2, 3, 5, 10, and 11) and includes thresholds for features 5, 10 and 11. The accuracy of the second decision rule on RTE3$_{Devmt}$

set was 71.50%, compared with an accuracy of 65.87 % on $RTE3_{Test.}$

Our results are consistent with previous RTE2 findings (Bar-Haim et al., 2006) where task performance varies with respect to the sentence source. Both rules had similar (poor) performance for information extraction (50.00 vs. 50.50%). Both rules had moderate performance for summarization (56.50 vs. 60.50%) and good performance for information retrieval (70.00 vs. 75.50%). The second decision rule constantly out-performed the first, with the largest increase of 11.5% in the question answering activity (65.50 vs. 77.00%).

Both decision rules lend themselves well to ranking sentences in decreasing order from the most to the least certain entailment. Average precision is calculated using that ranking and produces a perfect score when all sentence pairs that are entailments (+ve) are listed before all the sentence pairs that are not (-ve) (Voorhees & Harman., 1999). The average precision of the first and second decision rules was 58.97% and 60.96% respectively. The variation in precision also varied with respect to the sentence source (IE, IR, QA and SUM) of 48.52, 65.93, 72.38, and 56.04% for the first decision rule and 48.32, 72.71, 78.75 and 56.69% for the second decision rule.

## 4 Conclusions

Although most systems include both syntax and semantics to detect entailment and paraphrasing, our goal in this paper was to measure the impact of sentence structure alone. We developed two decision rules that each use features from a typed dependency grammar representation the hypothesis and test sentences. The first decision rule considers all features and the second considers only a sub-set of features, and adds thresholds to ensure that the system does not consider dubious matches. Thresholds for both rules were established using sentences in $RTE3_{Devmt}$ only. The second rule out-performed the first on $RTE3_{Test}$, both with respect to accuracy (60.50% vs. 65.87%) and average precision (58.97% vs. 60.96%).

These results are particularly encouraging given that our approach requires no background knowledge (other than the lexicon) and that this was the first time we participated in RTE. The results suggest that sentence structure alone can improve entailment prediction by between 9.25-14.62% alone, over the majority class baseline (51.52% in $RTE3_{Test}$) and they provided additional support to the growing body of evidence that sentence structure will continue to play a role in the accurate detection of textual entailments and paraphrasing.

## References


Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., et al. (2006). *The Second PASCAL Recognising Textual Entailment Challenge.* Venice, Italy.

Dagan, I., Glickman, O., & Magnini, B. (2005). *The PASCAL Recognising Textual Entailment Challenge.* Southampton, U.K.

de Marneffe, M.-C., MacCartney, B., Grenager, T., Cer, D., Rafferty, A., & Manning, C. D. (2006). *Learning to distinguish valid textual entailments*, In *The Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

Hickl, A., Williams, J., Bensley, J., Roberts, K., Rink, B., & Shi, Y. (2006). *Recognizing Textual Entailment with LCC's GROUNDHOG System* In *The Second PASCAL Recognising Textual Entailment Challenge,* Venice, Italy.

Klein, D., & Manning, C. D. (2002). *Fast Exact Inference with a Factored Model for Natural Language Parsing.* Paper presented at the Advances in Neural Information Processing Systems.

Lin, D. (1998). *Dependency-based Evaluation of MINIPAR.* In Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation, Granada,Spain.

Lin, D., & Pantel, P. (2001). Induction of semantic classes from natural language text. In *The 7th International conference on Knowledge discovery and Data Mining, San Francisco, CA.*

Litkowski, K. (2006). *Componential Analysis for Recognizing Textual Entailment.* In *The Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

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

National Library of Medicine. (2000). *The SPECIALIST Lexicon*, from www.nlm.nih.gov/pubs/factsheets/umlslex.html

Vanderwende, L., Coughlin, D., & Dolan, B. (2005). *What Syntax can Contribute in Entailment Task.* The PASCAL Challenges Workshop on Recognising Textual Entailment. Southhampton, UK.

Voorhees, E. M., & Harman., D. (1999). *Overview of the seventh text retrieval conference.* In *The Seventh Text REtrieval Conference* (TREC-7).