What Syntax can Contribute in Entailment Task

Lucy Vanderwende, Deborah Coughlin, Bill Dolan

Microsoft Research
Redmond, WA 98052
{lucyv; deborahc; billdol}@microsoft.com

Abstract

We describe our submission to the PASCAL Recognizing Textual Entailment Challenge which attempts to isolate the set of T-H pairs whose categorization can be accurately predicted based solely on syntactic cues. Two human annotators examined each pair, showing that a surprisingly large proportion of the data – 37% of the test items – can be handled with syntax alone, while adding information from a general-purpose thesaurus increases this to 49%.

1 Introduction

The data set made available by the PASCAL Recognizing Textual Entailment Challenge provides a great opportunity to focus on a very difficult task, determining whether one sentence (the hypothesis, H) is entailed by another (the text, T).

Our goal was to isolate the class of T-H pairs whose categorization can be accurately predicted based solely on syntactic cues. Human annotators made this judgment; we wanted to abstract away from the analysis errors that any specific parsing system would inevitably introduce. This work is part of a larger ablation study aimed at measuring the impact of various NLP components on entailment and paraphrase.

We have chosen to provide a partial submission that addresses the following question: what proportion of the entailments in the PASCAL test set could be solved using a robust parser? We are encouraged that other entrants chose to focus on different baselines, specifically those involving lexical matching and edit distance. Collectively, these baselines should establish what the minimal system requirements might be for addressing the textual entailment task.

2 Details of MSR submission

Various parsers providing constituent level analysis are now available to the research community, and state-of-the-art parsers have reported accuracy of between 89% and 90.1% F-measure (Collins and Duffy, 2002, Henderson 2004, and see Ringger et al., 2004, for a non-treebank parser). There are also efforts to produce parsers that assign argument structure (Gildea and Jurafsky, 2002, and for example, Hacioglu et al., 2004). With these developments, we feel that syntax can be defined broadly to include such phenomena as argument assignment, intra-sentential pronoun anaphora resolution, and a set of alternations to establish equivalence on structural grounds.

In order to establish a baseline for the entailment task that reflects what an idealized parser could accomplish, regardless of what any specific parser can do, we annotated the test set as follows. Two human annotators evaluated each T-H pair, deciding whether the entailment was:

- True by Syntax,
- False by Syntax,
- Not Syntax,
- Can't Decide

Additionally, we allowed the annotators to indicate whether recourse to information in a general purpose thesaurus entry would allow a pair to be judged True or False. Both annotators were skilled linguists, and could be expected to determine what an idealized syntactic parser could accomplish. We should note at this point that it could prove impossible to automate the judgment process described in this paper; the rules-of-thumb used by the annotators to make True of False judgments may turn out to be incompatible with an operational system.

We found that 37% of the test items can be handled by syntax, broadly defined; 49% of the test items can be handled by syntax plus a general purpose thesaurus. The results of this experiment are summarized in table 1:

	Without thesaurus	Using thesaurus
True	78 (10%)	147 (18%)
False	217 (27%)	244 (31%)
Not syntax	505 (63%)	409 (51%)

Table 1: Summary of MSR partial submission; Run1 is without thesaurus, Run2 is with thesaurus

Overall, inter-annotator agreement was 72%. Where there were disagreements, the annotators jointly decided which judgment was most appropriate in order to annotate all test items. Of the disagreements, 60% were between False and Not-Syntax, and 25% between True and Not-Syntax; the remainder of the differences were either annotation errors or where one or both chose Can't Decide. This confirms our anecdotal experience that it is easier to decide when syntax can be expected to return True, and that the annotators were uncertain when to assign False. In some cases, there are good syntactic clues for assigning False, which is why we designed the evaluation to force a choice between True, False, and Not-Syntax. But in many cases, it is the absence of syntactic equivalence or parallelism rather than a violation that results in a judgment of False, and most of the disagreements centered on these cases.

3 Results of Partial Submission

Our test results are not comparable to those of other systems, since obviously, our runs were produced by human annotators. In this section, we only want to briefly call attention to those test items where there was a discrepancy between our adjudicated human annotation and those provided as gold standard. It is worth mentioning that we believe the task is well-defined; for the 295 test items returned in Run1 of our submission, 284 matched the judgment provided as gold standard, so that our inter-annotator agreement with the test set is 96%.

In Run1 (using an idealized parser, but no thesaurus), there were 11 discrepancies. Of the 3 cases where we judged the test item to be True but the gold standard for the item is False, one is clearly an annotation error (despite having two annotators!) and two are examples of strict inclusion, which we allowed as entailments but the data set does not (test items 1839 and 2077); see (1).

(1) (pair id="2077", value="FALSE", task="QA")
<T> They are made from the dust of four of Jupiter's tiniest moons.
<H> Jupiter has four moons.

More difficult to characterize as a group are the 8 cases where we judged the test item to be False but the gold standard for the item is True (although 5/8 are from the QA section) The test items in question are: 1335, 1472, 1487, 1553, 1584, 1586, 1634, and 1682. It does appear to us that more knowledge is needed to judge these items than simply what is provided in the Text and Hypothesis, and these items should be removed from the data set accordingly since pairs for which there was disagreement among the judges were discarded. Item 1634 is a representative example.

(2) (pair id="1634", value="TRUE", task="IE")
<T> William Leonard Jennings sobbed loudly as was charged with killing his 3-year-old son, Stephen, who was last seen alive on Dec. 12, 1962.
<H> William Leonard Jennings killed his 3-year-old son, Stephen.

4 Requirements for a syntax-based system

There are many examples where predicate-argument assignment will give clear evidence for the judgment. (3a) and (3b) provide a good illustration:

(3) <T> Latvia, for instance, is the lowest-ranked team in the field but defeated World Cup semifi-

nalist Turkey in a playoff to qualify for the final 16 of Euro 2004.

(3a) <H> Turkey is defeated by Latvia.
(pair id="1897", value="TRUE", task="IE")

(3b) <H> Latvia is defeated by Turkey.
(pair id="1896", value="FALSE", task="IE")

4.1 Syntactic Alternations

By far the most frequent alternation between Text and Hypothesis that a system needs to identify is an appositive construction promoted to main clause. This alternation accounted for approximately 24% of the data.

(4) (pair id="760", value="TRUE", task="CD")
<T> The Alameda Central, west of the Zocalo, was created in 1592.
<H> The Alameda Central is west of the Zocalo.

Examples of other alternations that need to be identified are: nominalization → tensed clause (*Schroeder's election → Shroeder was elected*), shown in (5), and finite → non-finite construction (*where he was surfing → while surfing*), shown in (6).

(5) (pair id="315", value="TRUE", task="IR")
<T> The debacle marked a new low in the erosion of the SPD's popularity, which began shortly after Mr Schroeder's election in 1998.
<H> Schroeder was elected in 1998.

(6) (pair id="1041", value="TRUE", task="RC")
<T> A 30-year-old man has been killed in a shark attack at a surfing beach near Perth in West Australia where he was surfing with four other people.
<H> A 30-year-old man was killed in a shark attack while surfing.

4.2 Establishing False Entailment

We found two main categories of T-H pairs that we judged to be False: False, where there was a violation of a syntactic nature, and False, where there was no syntactic structure shared by the T-H pair. Although we can annotate this by hand, we are unsure whether it would be possible to create a system to automatically detect the absence of syntactic overlap, though the main verb in the Hypothesis should be the initial area of focus.

Examples of judging False by violation of syntax are those in which the Subject and Verb align (with or without thesaurus), but the Object does not, as in (7):

(7) (pair id="103", value="FALSE", task="IR")
<T> The White House ignores Zinni's opposition to the Iraq War.
<H> White House ignores the threat of attack.

The following examples illustrate an absence of shared syntactic structure in the major argument positions. In (8), the entailment is judged False since *baby girl* is not the subject of any verb of *buying*, nor is *ambulance* the object of any verb of *buying*; additionally, there is no mention of *buying* in T at all. In (9), the entailment is judged False because there is no mention of *Douglas Hacking* in the Text, nor any mention of *physician*. While a system using lexical matching might well rule the second example False, there are enough lexical matches in the former that a system using syntax is likely required.

(8) (pair id="2179", value="FALSE", task="RC")
<T> An ambulance crew responding to an anonymous call found a 3-week-old baby girl in a rundown house Monday, two days after she was snatched from her mother at a Melbourne shopping mall.
<H> A baby girl bought an ambulance at a Melbourne shopping mall.

(9) (pair id="2169", value="FALSE", task="CD")
<T> Scott and Lance Hacking talked with their younger brother at the hospital July 24.
<H>Douglas and Scott Hacking are physicians.

5 Interesting "Not Syntax" Examples

The number of examples that can be handled using syntax, broadly defined, is significant, but more than 50% were judged to be outside the realm of syntax, even allowing for the use of a thesaurus. Some test items exhibited phrasal-level synonymy, which the annotators did not expect would be available in a general purpose thesaurus. Consider, *X bring together Y* and *Y participate in X* in (10):

(10) (pair id="287", value="TRUE", task="IR")

<T> The G8 summit, held June 8-10, brought together leaders of the world's major industrial democracies, including Canada, France, Germany, Italy, Japan, Russia, United Kingdom, European Union and United States.

<H>Canada, France, Germany, Italy, Japan, Russia, United Kingdom and European Union participated in the G8 summit.

There are some examples with apparent alternation, but the alternation cannot be supported by syntax. Consider *three-day* and *last three days* in the following example:

(11) (pair id="294", value="TRUE", task="IR")

<T> The three-day G8 summit will take place in Scotland.

<H> The G8 summit will last three days.

In other cases, the annotators considered that there were too many alternations and thesaurus replacements necessary to confidently say that syntax could be used. Consider the following example, where *more than half* has to align with *many*, *saying* aligns with *thinking*, and *not worth fighting* aligns with *necessary*.

(12) (pair id="306", value="TRUE", task="IR")

<T> The poll, for the first time, has more than half of Americans, 52 percent, saying the war in Iraq was not worth fighting.

<H> Many Americans don't think the war in Iraq was necessary.

6 Discussion and Conclusion

Our goal is to contribute a baseline consisting of a system which uses an idealized parser, broadly defined, that can detect alternations, and optionally has access to a general purpose thesaurus. In order to explore what is possible, in principle, we used two human annotators and resolved their disagreements to produce a partial submission. It is interesting to note that the task is well-defined; for the 295 test items returned in our submission (without thesaurus), 284 matched the judgment provided as gold standard, so that our inter-annotator agreement is 96%.

A syntax-based system can account for 37% of the test items, and, with the addition of information from a general purpose thesaurus, 49%. This finding is promising, though we expect the numbers to decrease subject to an implementation with a real-world parser and set of matching rules. We also are keen to compare our baseline results with those obtained by the systems using lexical matching and edit distance, as we expect that some of the items that can be handled by syntax alone could also be accounted for by these simpler methods.

We hope that the challenge workshop is well served by offering these baselines, as it is clear to us that more than half of the test items represent an opportunity to work on very interesting entailment and paraphrase problems.

Acknowledgements

The authors wish to thank the organizers of the RTE Challenge and PASCAL for making creating this dataset and making it available. The authors also thank the Butler Hill Group for designing and maintaining the annotation environment.

References

Michael Collins and Nigel Duffy. 2002. New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron. *Proceedings of ACL 2002*, Philadelphia, PA.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3):245-288.

Kadri Hacioglu, Sameer Pradhan, Wayne Ward, James H. Martin, and Daniel Jurafsky, 2004. Semantic Role Labeling by Tagging Syntactic Chunks. *Proceedings of the Eighth Conference on Natural Language Learning (CONLL-2004)*, Boston, MA, May 6-7.

James Henderson. 2004. Discriminative training of a neural network statistical parser. *Proceedings of ACL 2004*, Barcelona, Spain.

Eric Ringger, Robert C. Moore, Eugene Charniak, Lucy Vanderwende, and Hisami Suzuki. 2004. Using the Penn Treebank to Evaluate Non-Treebank Parsers. *Proceedings of the 2004 Language Resources and Evaluation Conference (LREC)*. Lisbon, Portugal.