Recognizing Textual Entailment at RTE4 with CERES

Demetrios G. Glinos
Science Applications International Corporation
Orlando, Florida 32819
United States
glinosd@saic.com

Abstract

This paper reports on the performance of a new system for recognizing textual entailment. We present a brief overview of the CERES system and summarize its semantic alignment method for determining entailment. Then, we describe the experiments that were performed, both official and subsequently, against the RTE4 test dataset. We also report on comparison runs against the previous RTE3 development and test datasets.

1 Introduction

Recognizing textual entailment (RTE) involves the determination whether one piece of text, the hypothesis, logically follows from another piece of text. The ability to make such determinations is considered essential for a number of natural language processing tasks, such as question answering, summarization, information extraction, and machine translation (Giampiccolo et al., 2007).

The RTE workshop series[1] has served to focus research in this area by providing test data and a forum for evaluating entailment systems. For 2008, RTE4 has continued the three-way classification task that was piloted in RTE3, in which systems were asked to determine whether the hypothesis is entailed by the text, or contradicted by it, or whether neither result can be ascribed.

The CERES ("Concept Extraction and Reasoning System") system was developed for making such three-way determinations. It implements a semantic alignment framework in which both the hypothesis and text are processed into separate commitment sets of semantically role-tagged propositions, which are then matched using a semantic role-based alignment algorithm. The commitment sets represent the propositions that are expressly asserted and also implied by the input text and hypothesis. The basic idea behind the approach is that if these commitment sets are sufficiently complete, then entailment should be determinable by straightforward pattern matching, or at most by small-step inferencing from what may be found in the commitment sets.

The sections that follow present a brief overview of the CERES system, a description of the experiments that were performed, and the results obtained. In the last section, we offer our conclusions concerning the semantic alignment approach.

2 CERES System Overview

The principal functional components of the CERES system are shown in Figure 1. As shown, CERES employs a cascade of six components which operate as a pipeline to produce the hypothesis and text commitment sets. This knowledge base is then used by the entailment subsystem as the basis for entailment determinations. Each of these subsystems is discussed below.

2.1 Commitment Set Creation

The input text and hypothesis are transformed into their respective commitment sets by multi-stage pipeline of functions, as follows:

Syntactic Parsing: The inputs are separated into sentences and run through the Charniak parser (Charniak, 2000) to generate syntactically tagged parse trees.

Syntactic Analysis: The parse trees are decomposed into chunks and assigned syntactic roles; separate propositions are created for appositions,

[1] Originally sponsored by the PASCAL Network (see www.pascal-network.org/Challenges/RTE3), RTE is currently a track in the Text Analysis Conference (see www.nist.gov/tac/tracks/2008/rte/).

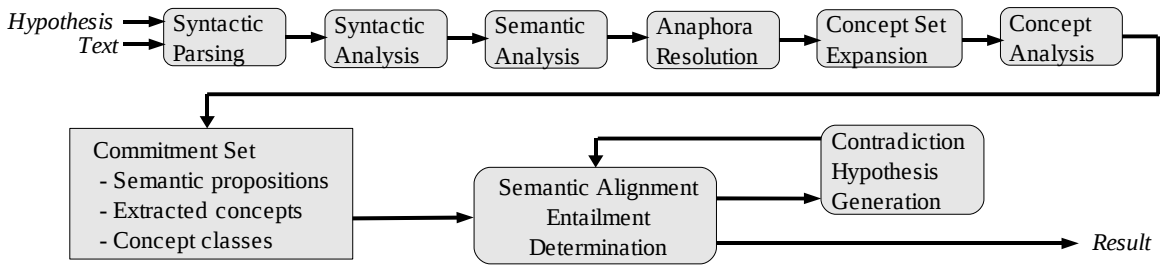

Figure 1. CERES Functional Components

parentheticals, relative and subordinate clauses, and similar constructs.

Semantic Analysis: Semantic arguments and adjuncts are identified; all propositions are expressed as semantically role-tagged logical structures using PropBank categories (Palmer, 2005).

Anaphora Resolution: Concepts are extracted from the semantic propositions; and pronominal and some elliptical references are resolved, where possible.

Commitment Set Expansion: Semantic propositions are manipulated to produce additional propositions representing what may be reasonably inferred from the input.

Concept Analysis: Concepts are analyzed to produce additional propositions; equivalent concepts are associated with one another in equivalence classes.

2.2 Entailment Determination

CERES employs a semantic alignment pattern matching algorithm for determining entailment. It uses the propositions in the hypothesis commitment set as templates to be matched by those in the commitment set generated from the companion text.

For each hypothesis proposition, the algorithm searches for a text proposition whose role set is compatible with those of the hypothesis. All hypothesis roles must be matched separately. Additional text proposition roles are ignored.

Individual roles are matched using a string matching algorithm which requires that each nonstopword in the hypothesis phrase for the role be matched in the corresponding text phrase. A successful match will be found if the words match exactly, or if they have common synonyms or hypernyms, as obtained from WordNet (Fellbaum, 1998). Equivalent concepts for both text and hypothesis phrases are also examined, if their respective concept equivalence classes are nonempty.

If any of the hypothesis propositions is successfully matched, then an affirmative entailment decision is reported. But if no match is found, then contradiction hypotheses are generated from the affirmative hypotheses, and an attempt is made to match those. If such a match is found, then a contradiction is reported. If neither type of match is found, then the algorithm reports that entailment is unknown for the input pair.

3 Experiments and Results

For RTE4, three official runs were submitted: a 2-way run ("CERES1_2W") and two 3-way runs ("CERES1_3W" and "CERES2_3W"). The second 3-way run differed from the first in that tighter contradiction hypothesis generation was employed, resulting in fewer contradiction findings.

Post-submission, we also performed a few experiments for comparison. First, we made a number of minor bug fixes and corrections to the software, but did not change its architecture. We also ran the final configuration against the previous RTE3 test and development 3-way datasets.

To place the results in context, we note first that the RTE4 test set consisted of 1,000 text-hypothesis pairs, of which 500 were instances of entailment, 150 were instances of contradiction, and 350 were instances of unknown entailment. Since CERES defaulted all results to unknown unless either entailment or contradiction were affirmatively determined, an accuracy value of .350 represents baseline performance. By comparison, the RTE3 test set contained 410 entailment pairs, 319 unknowns, and 72 contradictions.

Table 1 presents the key performance results for these experiments.

	2-way	3-way (1)	3-way (2)	Post	RTE3 Test	RTE3 Dev
2-way accuracy	.521	.526	.526	.575	.546	.596
3-way accuracy	-	.405	.416	.460	.471	.512
Precision	-	.708	.744	.806	.805	.847
Recall	-	.209	.206	.269	.249	.247
F-score	-	.444	.450	.538	.517	.570

Table 1. Experimental Results

The row for "3-way" accuracy represents the combined accuracy for all three possible result classes, while "2-way" accuracy represents the accuracy value where the outcomes for contradiction and unknowns are conflated.

Precision, recall, and F-score were computed only over entailed and contradiction results, with $\beta = 1/3$ to emphasize precision over recall, as was done for the RTE-3 three-way pilot task.

Viewed in the above context, we observe first that the results are consistent across all runs, with an expected minor improvement in performance for the post-submission bug fix run. We also observe that overall three-way accuracy is low, which is no doubt due to the low recall values. However, precision was uniformly high, above 70%. This is, apparently, a performance regime that we share with other logic-based systems (see MacCartney and Manning, 2007).

We note also that for all 3-way runs, the 2-way accuracy result is significantly higher than the 3-way result. We believe this is due the very low recall for our contradiction implementation. As explained in (de Marneffe, et al., 2008, at page 1041), the successful detection of contradiction requires the detection of "contradictions arising from the use of factive or modal words, structural and subtle lexical contrasts, as well as world knowledge (WK)," which our current implementation does not do.

Table 2 decomposes the precision results for the runs according to the four NLP applications areas from which the test set was drawn: Information Extraction (IE), Information Retrieval (IR), Question Answering (QA), and Summarization (SUM). For the RTE4 test set, there were 300 test pairs each for IE and IR, and 200 each for QA and SUM. For the RTE3 development and test sets, there were 200 test pairs for each type.

	3-way (1)	3-way (2)	Post	RTE3 Test	RTE3 Dev
IE	.727	.731	.736	.757	.738
IR	.722	.754	.866	.808	.857
QA	.657	.742	.824	.923	.945
SUM	.707	.743	.771	.733	.879
Combined	.708	.744	.806	.805	.847

Table 2. Precision Results

As shown in the table, precision performance was approximately the same across all NLP task areas, with the somewhat lower QA precision for the first RTE4 3-way run apparently having been corrected by the tighter contradiction generation used in the second run, which was the only difference between the two runs. The table also shows somewhat higher precision for both of the RTE3 QA runs, indicating that the QA pairs in those datasets may have been "easier" to process, although this has not been investigated in detail.

4 Conclusions and Recommendations

The results above demonstrate the viability of the semantic alignment approach for determining textual entailment. In particular, we have shown that the approach can produce high precision entailment determinations in all tested NLP task areas. However, in line with other logic based approaches, the current implementation suffers from low recall, which raises the question whether this approach can be extended to achieve higher levels of recall.

We believe that further investigation of this approach is warranted, not only because the current implementation relies on extensible heuristics for extending the commitment set, but also because the underlying parsing and syntactic analysis technologies, upon whose outputs the heuristics operate, are constantly improving.

The results also show a clear need to improve contradiction determination. It is evident that the current explicit contradiction hypothesis approach needs to be augmented by methods to infer contradiction from the overall context of a passage.

References

Eugene Charniak. 2000. A Maximum-Entropy-Inspired Parser. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, pages 132-139. Seattle, Washington. Association for Computing Machinery.

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding Contradictions in Text. In *Proceedings of ACL-08: HLT*, pages 1039-1047. Columbus, Ohio, USA. Association for Computational Linguistics.

Christiane Fellbaum, Ed. 1998. *WordNet: An Electronic Lexical Database.* The MIT Press, Cambridge, Massachusetts.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The Third PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the Workshop on Textual Entailment and Paraphrasing*, pages 1-9. June 28-29, 2007. Prague, Czech Republic. Association for Computational Linguistics.

Bill MacCartney and Christopher D. Manning. 2007. Natural Logic for Textual Inference. In *Proceedings of the Workshop on Textual Entailment and Paraphrasing*, pages 193-200. June 28-29, 2007. Prague, Czech Republic. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1): 71-106.