Workshop: Application of LCC's GROUNDHOG System for RTE-4

Jeremy Bensley and Andrew Hickl
Language Computer Corporation
1701 North Collins Boulevard
Richardson, Texas 75080 USA
{jeremy, andy}@languagecomputer.com

Abstract

This paper provides a brief description of the system for recognizing textual entailment (RTE) Language Computer Corporation (LCC) used in the 2008 TAC RTE-4 Evaluation. Our RTE-4 work follows our previous work (Hickl and Bensley, 2007; Bensley and Hickl, 2008; Hickl, 2008) in using a pipeline of lightweight, largely statistical systems for commitment extraction, lexical alignment, and entailment classification in order to estimate the likelihood that a *t* includes the linguistic content sufficient to textually entail a *h*. Our system continues to provide promising results: on the binary entailment task, our system correctly classified more than 74% of *t*-*h* pairs correctly.

1 Introduction

This paper provides a brief description of the system for recognizing textual entailment (RTE) Language Computer Corporation (LCC) used in the 2008 TAC RTE-4 Evaluation.

This year's evaluation marks the third consecutive year that LCC has participated in a PASCAL or TAC-sponsored RTE evaluation. While we have experimented with different approaches designed to increase the amount of lexicosemantic knowledge available to an RTE system, our basic approach to the problem of RTE has remained largely unchanged over the past three years. Each of our RTE systems (RTE-2 (Hickl et al., 2006), RTE-3 (Hickl and Bensley, 2007), and this work) has used a lightweight statistical framework in order to estimate the likelihood that a *t* textually entails an *h*. Under this framework, we assume that hypotheses which share more lexical and semantic features with one of the propositions inferable from a text will be more likely to be textually entailed; in contrast, hypotheses which differ markedly, or include potentially contradictory (Harabagiu et al., 2006) features, will be less likely to represent a valid instance of TE.

Our RTE-2 system (Hickl et al., 2006) followed influential work by (Marsi and Krahmer, 2005; Haghighi et al., 2005; Raina et al., 2005) in using features derived from a graph-matching algorithm to inform a Maximum Entropy-based entailment classifier. In order to increase the amount of lexicosemantic information available to this system, we followed (Burger and Ferro, 2005) in using a heuristic-based approach in order to generate 450,000 additional entailment pairs which could be used to train an entailment classifier. In addition to these training examples, we also used a Web-based paraphrase generation algorithm (based on (Barzilay and Lee, 2003)) to create "rough" paraphrases of 2- or 3-place predicates included in either the *t* or *h*.

Our RTE-3 system (Hickl and Bensley, 2007) used a lightweight information extraction framework in order to enumerate all of the possible propositions that could be inferred from a *t*-*h* pair. Once these commitments had been decoded from an entailment pair, we showed that RTE could be performed with greater than 80% accuracy – even without access to additional sources of linguistic knowledge. (In a departure from our RTE-2 work, our RTE-3 system did not use any additional training examples or automatically-generated paraphrases in its entailment estimation.)

This year, we used an implementation of the RTE system described in (Hickl, 2008) to participate in RTE-4. As such, the system description we present in this (notebook) paper will share many similarities with that work. We have been careful to point out where the system deployed in RTE-4 differs from the system developed for that previous work. This system is very similar to our RTE-3 work in its use of a commitment-based extraction framework; a discussion of how our current system differs from this previous work is provided in (Hickl, 2008).

The rest of this paper is organized in the following way. Section 2 provides a sketch of the system we used in the PASCAL RTE-4 Challenge, while Section 3 discusses results from this year's evaluation, and Section 4 provides our conclusions.

2 System Overview

In our work, we use our own implementation of the lightweight method for RTE described in (Hickl and Bensley, 2007) in order to compute the likelihood that an edge ($x \rightarrow y$) holds between a pair of nodes (x, y).

Algorithm 1 Extracting Discourse Commitments

```
Input: An underlying argument structure $S$
Output: A syntactic tree $Q$ corresponding to a well-formed natural language question
for all Candidate answer $c$ identified in $\{a_i, ..., a_m\} \in S$ do
    Adjoin($c$,$q_0$)
    Associate $c$ with the root node ($q_0$) of syntactic tree $Q$ corresponding to the generated question
    for all Non-terminal nodes $\{q_1, ..., q_n\}$ in $Q$ do
        for all Grammar rules $\{r_i, ..., r_m\}$ which apply to $q$ do
            Generate structure $K : \{q'_i, ..., q'_n\}$ described by $r$
            Determine $p(K)$
            for all Grammar rules $\{r'_i, ..., r'_m\}$ which apply to $K$ lower in $Q$ do
                Generate the structure $K' : \{q''_i, ..., q''_n\}$ described by $r'$
                Determine $p(K')$
            end for
            if $p(K) \leq p(K')$ then
                Discard($K$)
            end if
        end for
        Adjoin($K$,$Q$) which argmax $p(K)$
    end for
end for
```

First described in (Dagan et al., 2005), the task of recognizing textual entailment (RTE) requires systems to determine whether the meaning of a short text passage can be conventionally inferred from the meaning of some longer text passage. While the recognition of textual inference (such as TE) has traditionally been addressed using formal reasoning methods (such as automatic theorem proving (Tatu et al., 2006), model checking, or model building (Blackburn and Bos, 2005)), a considerable amount of recent work conducted as part of the PASCAL Recognizing Textual Entailment Challenges (Dagan et al., 2005; Bar-Haim et al., 2006; Giampiccolo et al., 2007) have demonstrated the viability of relatively "shallow", statistical learning-based approaches to RTE. The architecture of our system for recognizing textual entailment is presented in Figure 1.

We follow (Hickl and Bensley, 2007) in using adopting an approach to RTE which explicitly considers the set of *discourse commitments* that are derivable from the textual content of a pair of texts. In our work, we used the probabilistic finite-state transducer (FST)-based information extraction framework described in (Eisner, 2002) to extract commitments from four different types of syntactic constructions, including: (1) supplemental expressions (Huddleston and Pullum, 2002) (such as as-clauses, non-restrictive relative clauses, nominal appositives, parenthetical adverbs, and epithets), (2) coordination, (3) subordination, and (4) possessives. Commitments were extracted using a series of weighted regular expressions that were created by hand; weights were learned for each regular expression $r \in R$ using our implementation of (Eisner, 2002). After each candidate commitment was processed by the FST, each commitment was then resubmitted to the FST for additional round(s) of extraction until no additional commitments could be extracted from the input string.

Bobby Zamora scored in the 57th minute Monday
to lift West Ham over Preston North End.

Commitments

1. Bobby Zamora scored in the 57th minute.
2. Bobby Zamora scored [on] Monday.
3. Bobby Zamora lifted West Ham over Preston North End.
4. Bobby Zamora [is a member of] West Ham.
5. West Ham [is a sports team].
6. West Ham was lifted over Preston North End.

Figure 2: Sample Commitments

We then expand the total number of commitments

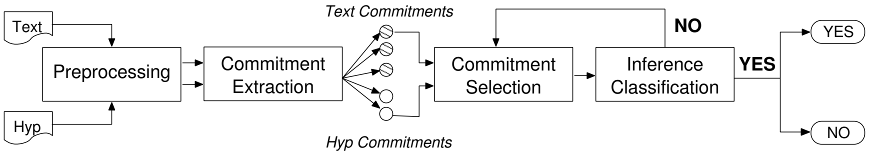

Figure 1: Architecture for Recognizing Textual Entailment.

available from each sentence using the lightweight approach to paraphrasing described in Section 2. Paraphrases are generated by first selecting a pair of arguments from each commitment; these arguments are then used to retrieve a set of sentences containing both arguments and supplied to a clustering algorithm in order to identify the constructions which represent the best possible paraphrases of the commitment.

Semantic dependencies identified by a PropBank-based semantic parser were then used to convert each commitment into a dependency graph, where nodes were associated with individual tokens (or phrases identified by a tokenizer) and edges were associated with the set of syntactic or semantic dependencies that link pairs of nodes. We then used a variant of the maximum weighted matching approach introduced in (Taskar et al., 2005) in order to identify the commitments in a $\{C_u\}$ which represent the best possible alignment for each commitment in $\{C_v\}$.

We used the *reciprocal best-hit method* (Mushegian and Koonin, 2005) in order to select pairs of aligned commitments from {C} to be considered in the recognition of textual entailment. Under this method, a pair of commitments (c_u, c_v) was considered iff c_u was found in the top-n alignments identified c_v and iff c_v was also found in the top-n alignments identified for c_u.[1]

Once the best set of commitment alignments $(\{(c_{u_1}, c_{v_1}), ..., (c_{u_n}, c_{v_n})\})$ have been identified for a pair of nodes, we use a decision tree classifier in order to estimate the likelihood (expressed as the confidence of classifying a pair of commitments as either a positive or negative instance of textual entailment) that a commitment from u textually entails a commitment derived from v. We learn this classifier using a set of linguistic features analogous to those described in several previous approaches, including (Hickl and Bensley, 2007; Marsi et al., 2007; MacCartney et al., 2006; Zanzotto et al., 2006). In our current model, we assume that the content of a node u textually entails a node v iff there exists at least some $c \in C_u$ such that $c \models_{te} v$.

3 Experiments and Results

We submitted one ranked run in our official submission for this year's evaluation. Our results from the RTE-4 Test Set are presented in Table 1.

	IE	IR	QA	SUM	Total
Accuracy	0.773	0.76	0.72	0.71	0.746
Average Precision	0.779	0.797	0.717	0.717	0.7419

Table 1: RTE-4 Results per Task.

An in-depth analysis of these results, plus an evaluation of a second post-hoc experiment will be provided in the final camera-ready version of this paper.

4 Conclusions

This paper provided a formal description of the RTE system Language Computer Corporation used in its participation in the TAC RTE-4 evaluation. Our primary objective for participating in this year's evaluation was to test the robustness of a mature system (Hickl and Bensley, 2007; Hickl, 2008) on a new RTE dataset. We expect that our participation in this year's evaluation will enable us to identify key technology areas that should be targeted for further research and development. We found that our system continued to perform well on this year's Test Set, despite any major modifications: our RTE-4 system correctly classified 74.6% of *t-h* pairs as part of the binary entailment task.

In future work, we plan to adapt our current classification-based infrastructure to address multi-way entailment tasks, such as a ternary classification (YES, NO, UNKNOWN) or any *n*-ary classification task that may be considered in future evaluations. We anticipate that expanding the number of classification outcomes will force us to rethink how

[1] Values for n were selected by optimizing the performance of the entailment classifier on our training set.

linguistic evidence is extracted – and marshalled – in support of an entailment judgment and may require a recasting of the commitment extraction framework which our current RTE system is based on.

5 Acknowledgments

This space is reserved for the acknowledgments which will be included in the final camera-ready version of this paper.

References

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The Second PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the Second PASCAL Challenges Workshop*.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *HLT-NAACL*.

Jeremy Bensley and Andrew Hickl. 2008. Unsupervised Resource Creation for Textual Inference Applications. In *Proceedings of the Sixth LREC Conference 2008 (LREC 2008)*, Marrakech, Morocco, May.

P. Blackburn and J. Bos. 2005. *Representation and Inference for Natural Language*. CSLI.

John Burger and Lisa Ferro. 2005. Generating an Entailment Corpus from News Headlines. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 49–54.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL Recognizing Textual Entailment Challenge. In *Proceedings of the PASCAL Challenges Workshop*.

Jason Eisner. 2002. Parameter estimation for probabilistic finite-state transducers. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1–8, Philadelphia, July.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague, June. Association for Computational Linguistics.

Aria Haghighi, Andrew Ng, and Christopher Manning. 2005. Robust textual inference via graph matching. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 387–394.

Sanda Harabagiu, Andrew Hickl, and Finley Lacatusu. 2006. Negation, Contrast, and Contradiction in Text Processing. In *Proceedings of AAAI*, Boston, MA.

Andrew Hickl and Jeremy Bensley. 2007. A discourse commitment-based framework for recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 171–176, Prague, June. Association for Computational Linguistics.

Andrew Hickl, John Williams, Jeremy Bensley, Kirk Roberts, Bryan Rink, and Ying Shi. 2006. Recognizing Textual Entailment with LCC's Groundhog System. In *Proceedings of the Second PASCAL Challenges Workshop*.

Andrew Hickl. 2008. Using Discourse Commitments to Recognize Textual Entailment. In *Proceedings of the 22nd COLING Conference*, Manchester, UK.

Rodney Huddleston and Geoffrey Pullum, editors, 2002. *The Cambridge Grammar of the English Language*. Cambridge-University Press.

Bill MacCartney, Trond Grenager, Marie-Catherine de Marneffe, Daniel Cer, and Christopher D. Manning. 2006. Learning to recognize features of valid textual entailments. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 41–48, New York City, USA, June. Association for Computational Linguistics.

Erwin Marsi and Emiel Krahmer. 2005. Classification of semantic relations by humans and machines. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 1–6.

Erwin Marsi, Emiel Krahmer, and Wauter Bosma. 2007. Dependency-based paraphrasing for recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 83–88, Prague, June. Association for Computational Linguistics.

Arcady Mushegian and Eugene Koonin. 2005. A minimal gene set for cellular life derived by compraison of complete bacterial genomes. In *Proceedings of the National Academies of Science*, volume 93, pages 10268–10273.

Rajat Raina, Andrew Y. Ng, and Chris Manning. 2005. Robust textual inference via learning and abductive reasoning. In *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI)*.

Ben Taskar, Simone Lacoste-Julien, and Dan Klein. 2005. A discriminative matching approach to word alignment. In *Proceedings of Human Language Technology Conference and Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, Canada.

Marta Tatu, Brandon Iles, John Slavick, Adrian Novischi, and Dan Moldovan. 2006. COGEX at the Second Recognizing Textual Entailment Challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

F. Zanzotto, A. Moschitti, M. Pennacchiotti, and M. Pazienza. 2006. Learning textual entailment from examples. In *Proceedings of the Second PASCAL Challenges Workshop*.