Using Discourse Commitments to Recognize Textual Entailment

Andrew Hickl
Language Computer Corporation
1701 North Collins Boulevard Suite 2000
Richardson, Texas 75080 USA
andy@languagecomputer.com

Abstract

In this paper, we introduce a new framework for recognizing textual entailment (RTE) which depends on extraction of the set of publicly-held beliefs – known as *discourse commitments* – that can be ascribed to the author of a text (*t*) or a hypothesis (*h*). We show that once a set of commitments have been extracted from a *t*-*h* pair, the task of recognizing textual entailment is reduced to the identification of the commitments from a *t* which support the inference of the *h*. Our system correctly identified entailment relationships in more than 80% of *t*-*h* pairs taken from all three of the previous PASCAL RTE Challenges, without the need for additional sources of training data.

1 Introduction

Systems participating in the PASCAL Recognizing Textual Entailment (RTE) Challenges (Dagan et al., 2005) have successfully employed a variety of "shallow" techniques in order to recognize instances of textual entailment, including methods based on: (1) sets of heuristics (Vanderwende et al., 2006), (2) measures of term overlap (Jijkoun and de Rijke, 2005) (or other measures of semantic "relatedness" (Glickman et al., 2005), (3) the alignment of graphs created from syntactic or semantic dependencies (Haghighi et al., 2005), or (4) statistical classifiers which leverage a wide range of features, including the output of paraphrase generation (Hickl et al., 2006), inference rule generation (Szpektor et al., 2007), or model building systems (Bos and Markert, 2006).

While these relatively "shallow" approaches have shown much promise in RTE for entailment pairs where the text and hypothesis remain short, we expect that performance of these types of systems will ultimately degrade as longer and more syntactically complex entailment pairs are considered. For example, given a "short" t-h pair (as in (1)), we might expect that a feature-based comparison of the t and the h would be sufficient to identify that the t textually entailed the h.

(1) Short t-h Pair

a. **Text:** Mack Sennett was involved in the production of "The Extra Girl".

b. **Hypothesis:** "The Extra Girl" was produced by Sennett.

The additional information included in a longer t (like the one in (2)) can make for a much more challenging entailment computation. While the evidence supporting the h is included in the t, systems must be able to establish that (1) *Mack Sennett* was involved in producing a *Mabel Normand vehicle* and (2) that *"The Extra Girl"* and the *Mabel Normand vehicle* refer to the same film.

(2) Long t-h Pair

a. **Text:** "The Extra Girl" (1923) is the story of a small-town girl, Sue Graham (played by Mabel Normand) who comes to Hollywood to be in the pictures. This Mabel Normand vehicle, produced by Mack Sennett, followed earlier films about the film industry and also paved the way for later films about Hollywood, such as King Vidor's "Show People" (1928).

b. **Hypothesis:** "The Extra Girl" was produced by Sennett.

In order to remain effective as texts get longer, we believe that RTE systems will need to employ techniques that will enable them to enumerate the

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 337–344
Manchester, August 2008

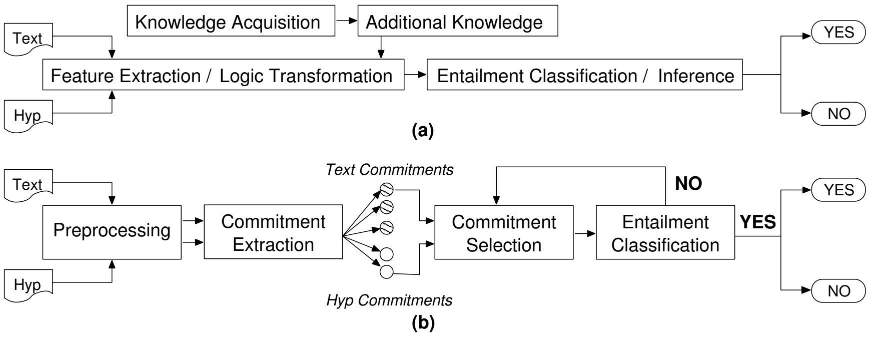


Figure 1: Two Architectures of RTE Systems.

set of propositions which are inferable from a text-hypothesis pair.

We introduce a new framework for recognizing textual entailment which depends on extraction of a subset of the publicly-held beliefs – or *discourse commitments* – available from the linguistic meaning of a text or hypothesis. We show that once even a small set of discourse commitments have been extracted from a text-hypothesis pair, the task of RTE can be reduced to the identification of the one (or more) commitments from the t which are most likely to support the inference of each commitment extracted from the h.

We have found that a commitment-based approach to RTE provides state-of-the-art results on the PASCAL RTE task even when large external knowledge resources are not available. While our approach does depend on a set of specially-tailored heuristics which makes it possible to enumerate some of the commitments from a *t-h* pair, we show that reasonably high levels of performance are possible (as much as 84.9% accuracy), given even a small number of extractors.

The rest of this paper is organized in the following way. Section 2 describes the organization of most current statistical systems for RTE, while Sections 3, 4, and 5 describe details of the algorithms we have developed for the RTE system we discuss in this paper. Section 6 presents our experimental results, and Section 7 presents our conclusions.

2 Recognizing Textual Entailment

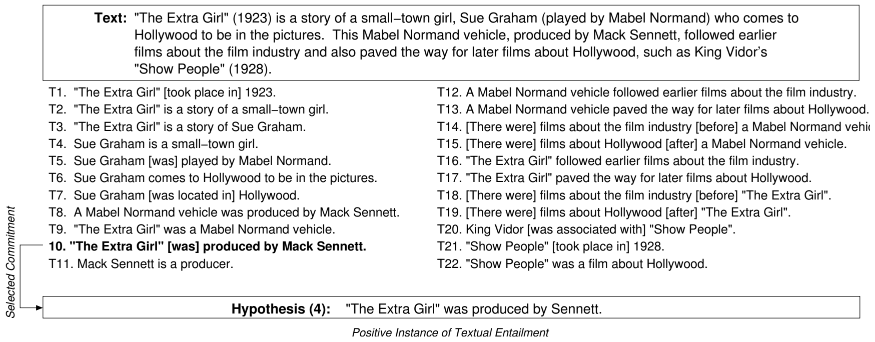
Recognizing whether the information expressed in a h can be inferred from the information expressed in a t can be cast either as (1) a classification problem or (2) a formal textual inference problem, performed either by theorem proving or model checking. While these approaches apply radically different solutions to the same problem, both methods involve the translation of natural language into some sort of suitable meaning representation, such as real-valued features (in the case of classification), or axioms or models (in the case of formal methods). We argue that performing this translation necessarily requires systems to acquire forms of (linguistic and/or real-world) knowledge which may not be derivable from the surface form of a t or h. (See Figure 1a for an illustration of the architecture of a prototypical RTE system.)

In order to acquire forms of linguistic knowledge for RTE, we have developed a novel framework which depends on the extraction of *discourse commitments* from a text-hypothesis pair. Following (Gunlogson, 2001; Stalnaker, 1979), we assume discourse commitments represent the set of propositions which can necessarily be inferred to be true given a conventional reading of a text. (Figure 2 lists the set of commitments that were extracted from a *t-h* pair included in the PASCAL RTE-3 Test Set.[1])

Formally, we assume that given a commitment set $\{c_t\}$ consisting of the set of discourse commitments inferable from a text t and a hypothesis h, we define the task of RTE as a search for the commitment $c \in \{c_t\}$ which maximizes the likelihood that c textually entails h.

In our architecture (illustrated in Figure 1b), discourse commitments are first extracted from both the t and the h using the approach described in Section 3. Once commitment sets have been extracted for the t and the h, we then use a *commitment selection* module (described in Section 4) in order to perform a term-based alignment of each commitment extracted from the t against each commitment extracted from the h. The top-ranked pair of commitments (c_{t_i}, c_{h_i}) is then sent to an

[1] Under our approach, commitments are extracted from both the t and the h. Our system failed to extract any commitments from the h used in this example, however.



Text: "The Extra Girl" (1923) is a story of a small–town girl, Sue Graham (played by Mabel Normand) who comes to Hollywood to be in the pictures. This Mabel Normand vehicle, produced by Mack Sennett, followed earlier films about the film industry and also paved the way for later films about Hollywood, such as King Vidor's "Show People" (1928).

T1. "The Extra Girl" [took place in] 1923.
T2. "The Extra Girl" is a story of a small–town girl.
T3. "The Extra Girl" is a story of Sue Graham.
T4. Sue Graham is a small–town girl.
T5. Sue Graham [was] played by Mabel Normand.
T6. Sue Graham comes to Hollywood to be in the pictures.
T7. Sue Graham [was located in] Hollywood.
T8. A Mabel Normand vehicle was produced by Mack Sennett.
T9. "The Extra Girl" was a Mabel Normand vehicle.
10. "The Extra Girl" [was] produced by Mack Sennett.
T11. Mack Sennett is a producer.
T12. A Mabel Normand vehicle followed earlier films about the film industry.
T13. A Mabel Normand vehicle paved the way for later films about Hollywood.
T14. [There were] films about the film industry [before] a Mabel Normand vehi
T15. [There were] films about Hollywood [after] a Mabel Normand vehicle.
T16. "The Extra Girl" followed earlier films about the film industry.
T17. "The Extra Girl" paved the way for later films about Hollywood.
T18. [There were] films about the film industry [before] "The Extra Girl".
T19. [There were] films about Hollywood [after] "The Extra Girl".
T20. King Vidor [was associated with] "Show People".
T21. "Show People" [took place in] 1928.
T22. "Show People" was a film about Hollywood.

Hypothesis (4): "The Extra Girl" was produced by Sennett.

Figure 2: Text Commitments Extracted from Example 4 (RTE-3).

entailment computation module (described in Section 5), which estimates the likelihood that the selected c_{t_i} textually entails the c_{h_i} (and by extension, the likelihood that t textually entails h). Commitment pairs are considered in ranked order until a positive judgment is returned, or until no more commitments above a threshold remain.

3 Extracting Discourse Commitments from Natural Language Texts

Work in *semantic parsing* (Wong and Mooney, 2007; Zettlemoyer and Collins, 2005) has used statistical and symbolic techniques to convert natural language texts into a logical meaning representation (MR) which can be leveraged by formal reasoning systems. While this work has explored how output from syntactic parsers can be used to represent the meaning of a text independent of its actual surface form, these approaches have focused on the propositional semantics explicitly encoded by predicates and have not addressed other phenomena (such as conversational implicature or linguistic presupposition) – which are not encoded overtly in the syntax.

Our work focuses on how an approach based on lightweight extraction rules can be used to enumerating a subset of the discourse commitments that are inferable from a *t-h* pair. While heuristically "unpacking" all of commitments of a t or h may be (nearly) impossible, we believe that our work represents an important first step towards determining the relative value of these additional commitments for textual inference applications, such as RTE. Our commitment extraction algorithm is presented in Algorithm 1.

Commitments are extracted from each t and h using an implementation of the probabilistic finite-state transducer (FST)-based extraction framework described in (Eisner, 2002; Eisner, 2003). Given a syntactically and semantically-parsed input string, our system returns a series of output representations which can be mapped (given a set of generation heuristics) to natural language sentences which represent each of the individual commitments which can be extracted from that string. Commitments were extracted using a series of weighted regular expressions (which we present in the form of rules for convenience); 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, the natural language form of each returned commitment was then resubmitted to the FST for additional round(s) of extraction until no additional commitments could be extracted from the input string.

Algorithm 1 Extracting Discourse Commitments

```
Input: A set $S$ of sentences from the $t$ or $h$
Output: A set of discourse commitments
loop
    Pre-process the decompositions to identify lexical, syntactic, semantic, and discourse information
    Produce syntactic decompositions of the sentences in $S$
    Produce commitments of propositional content
    Produce commitments for supplemental expressions
    Extract predefined set of relations
    Perform coreference resolution for each commitment
    Generate paraphrases of each commitment
    Generate a natural-language string for each commitment
    if any of the generated strings is not in $S$ then
        Add the generated strings to $S$
    else
        return $S$
    end if
end loop
```

Text-hypothesis pairs are initially submitted to a *preprocessing* module which performed (1) part-of-speech tagging, (2) named entity recognition, (3) syntactic dependency parsing, (4) semantic dependency parsing, (5) normalized temporal expressions, and (6) coreference resolution.[2]

Pairs are then submitted to a *sentence decomposition* module, which uses a set of heuristics in order to transform complex sentences containing subordination, relative clauses, lists, and coordination into sets of well-formed simple sentences.

Propositional Content: In order to capture assertions encoded by predicates and predicate nominals, we used semantic dependency information output by a predicate-based semantic parser to generate "simplified" commitments for each possible combination of their optional and obligatory arguments. (Here, arguments assigned a PropBank-style role label of $\textsc{Arg}_m$ – or $\textsc{Arg}_2$ or higher – were considered to be optional arguments.)

Supplemental Expressions: Recent work by (Potts, 2005; Huddleston and Pullum, 2002) has demonstrated that the class of supplemental expressions – including appositives, *as*-clauses, parentheticals, parenthetical adverbs, non-restrictive relative clauses, and epithets – trigger conventional implicatures (CI) whose truth is necessarily presupposed, even if the truth conditions of a sentence are not satisfied. Rules to extract supplemental expressions were implemented in our weighted FST framework; generation heuristics were then used to create new sentences which specify the CI conveyed by the expression.

(3) "The Extra Girl" (1923) is a story of a small-town girl, Sue Graham (played by Mabel Normand) who comes to Hollywood to be in the pictures.

a. "The Extra Girl" [took place in] 1923.
b. "The Extra Girl" is the story of a small-town girl.
c. Sue Graham is a small-town girl.
d. Sue Graham [was] played by Mabel Normand.
e. Sue Graham [came] to Hollywood to be in the pictures.

Relation Extraction: We used an in-house, heuristic-based relation extraction system to recognize six types[3] of semantic relations between named entities, including: (1) *artifact* (e.g. OWNER-OF), (2) *general affiliation* (e.g. LOCATION-OF), (3) *organization affiliation* (e.g. EMPLOYEE-OF), (4) *part-whole*, (5) *social affiliation* (e.g. RELATED-TO), and (6) *physical location* (e.g. LOCATED-NEAR) relations.

(4) Sue Graham [came] to Hollywood to be in the pictures.

a. **location-of:** Sue Graham [was located in] Hollywood.
b. **location-of:** The pictures [were located in] Hollywood.

Coreference Resolution: We use our own implementation of (Ng, 2005) to resolve instances of pronominal and nominal coreference in order to expand the number of commitments available to the system. After a set of co-referential entity mentions were detected (e.g. *"The Extra Girl"*, *this Mabel Normand vehicle*), new commitments were generated from the existing set of commitments which incorporated each co-referential mention.

(5) **Coreference:** ("The Extra Girl",this Mabel Normand vehicle)

a. ["The Extra Girl"] [was] produced by Mack Sennett.
b. ["The Extra Girl"] followed earlier films about the film industry.
c. ["The Extra Girl"] also paved the way for later films about Hollywood.

Paraphrasing: We used a lightweight, knowledge-lean paraphrasing approach in order to expand the set of commitments considered by our system.

In order to identify other possible linguistic encodings for each commitment – without generating a large number of spurious paraphrases which could introduce errorful knowledge into a commitment set – we focused only on generating paraphrases of two-place predicates (i.e. predicates which encode a semantic dependency between two arguments).

The algorithm we use is presented in Algorithm 2. Under this method, a semantic parser (trained on the semantic dependencies in PropBank and NomBank) was used to identify pairs of arguments ($\langle a_i, a_j \rangle$) (where $i, j \in \{a_0, a_1, a_2, a_m\}$) from each c; each pair of arguments identified in c are then used to generate paraphrases from sets of sentences containing both a_0 and a_i.[4] The top 1000 sentences containing each pair of arguments were then retrieved from

[2] We used publicly-available software when possible to facilitate comparison with other researchers' work in this area. We used the *C&C Tools* (Curran et al., 2007) to perform part-of-speech tagging and parsing and a version of the alias-i *LingPipe* named entity recognizer in conjunction with Language Computer Corporation's own systems for syntactic and semantic dependency parsing, named entity recognition (*CiceroLite*), temporal normalization, and coreference resolution.

[3] These six types were selected because they performed at better than than 70% F-Measure on a sample of *t-h* pairs selected from the PASCAL RTE datasets. Other relation types with lesser performance were not used in our experiments.

[4] Arguments in PropBank and NomBank are assigned an index corresponding to their semantic role.

the WWW; sentences containing both arguments were then filtered and clustered into sets that were presumed to be likely paraphrases.

Algorithm 2 Paraphrase Clustering

```
Input: Pairs of arguments, $\langle a_i, a_j \rangle$, where $i, j \in \{a_0, a_1, a_2, a_m\}$
Output: Sets of paraphrased sentences, $\{s_{cl_i \ldots cl_n}\}$
for all pairs $\langle a_i, a_j \rangle$ do
    Retrieve 1000 $s$ containing $\langle a_i, a_j \rangle$ from WWW
    Compute token distance between $a_0$,$a_i$ ($\texttt{span}(a_i, a_j)$)
    Filter each $s$ with $2 \leq \texttt{span}(a_i, a_j) \leq 8$
    Complete-link cluster $\{s\}$ into clusters $\{cl\}$
    Filter $cl$ with size ($\texttt{size}(cl)$) $\leq 10$
    return All sentences $\{s_{cl_i \ldots cl_n}\}$
    Add all sentences $\{s_{cl_i \ldots cl_n}\}$ to commitment set $C$
end for
```

Parameters were computed using maximum likelihood estimation (and normalized to sum to 1), based on a linear interpolation of three measures of the "goodness" of p (as compared to the original input sentence, s).

$$para(p|s) = \lambda_{wn} para(p|s) + \lambda_{freq} para(p|s) + \lambda_{dist} para(p|s) \tag{1}$$

Each candidate paraphrase was then assigned a *paraphrase score* as in (Glickman and Dagan, 2005). The likelihood that a word w_p from a paraphrase was a valid paraphrase of a word from an original commitment w_c was computed as in (2), where $p(w_p)$ and $p(w_o)$ computed from the relative frequency of the occurrence of w_p and w_o in the set of clusters generated for c, and $p(k)$ was computed from the frequency of each "overlapping" term found in the paraphrase and the original c. The top 5 paraphrases generated for each c were then used to generate new versions of each commitment.

$$p_{para}(w_p|w_o) = p(w_p)p(w_o) \prod{}_{i=1}{}^{n} p(k_i) \tag{2}$$

4 Commitment Selection

Following Commitment Extraction, we used a lexical alignment technique first introduced in (Taskar et al., 2005b) in order to select the commitment extracted from t (henceforth, c_t) which represents the best alignment for each of the individual commitments extracted from h (henceforth, c_h).

We assume that the alignment of two discourse commitments can be cast as a maximum weighted matching problem in which each pair of words (t_i,h_j) in an commitment pair (c_t,c_h) is assigned a score $s_{ij}(t, h)$ corresponding to the likelihood that t_i is aligned to h_j.[5] As with (Taskar et al., 2005b), we use the large-margin structured prediction model introduced in (Taskar et al., 2005a) in order to compute a set of parameters w (computed with respect to a set of features f) which maximize the number of correct alignment predictions ($\bar{y}_i$) made given a set of training examples (x_i), as in Equation (3).

$$y_i = \arg\max{}_{\bar{y}_i \in Y} w^{\top} f(x_i, \bar{y}_i), \forall i \tag{3}$$

We used three sets of features in our model: (1) string features (including Levenshtein edit distance, string equality, and stemmed string equality), (2) lexico-semantic features (including WordNet Similarity (Pedersen et al., 2004)), and (3) word association features (computed using the Dice coefficient (Dice, 1945)[6]). Training data came from hand-annotated token alignments for each of the 800 entailment pairs included in the RTE-3 Development Set

Following alignment, we used the sum of the edge scores ($\sum{}_{i,j=1}{}^{n} s_{ij}(t_i, h_j)$) computed for each of the possible (c_t, c_h) pairs in order to search for the c_t which represented the *reciprocal best hit* (Mushegian and Koonin, 1996) of each c_h extracted from the hypothesis. This was performed by selecting a commitment pair (c_t, c_h) where c_t was the top-scoring alignment candidate for c_h and c_h was the top-scoring alignment candidate for c_t. If no reciprocal best-hit could be found for any of the commitments extracted from the h, the system automatically returned a TE judgment of NO.

5 Entailment Classification

We used a decision tree to estimate the likelihood that a commitment pair represented a valid instance of textual entailment. Confidence values associated with each leaf node (i.e. YES or NO) were normalized and used to rank examples for the official submission. Features were selected manually by performing ten-fold cross validation on the combined development sets from the three previous PASCAL RTE Challenges (2400 examples).

[5] In order to ensure that content from the h is reflected in the t, we assume that each word from the h is aligned to exactly one or zero words from the t.

[6] The Dice coefficient was computed as $Dice(i) = \frac{2C_{th}(i)}{C_t(i)C_h(i)}$, where C_{th} is equal to the number of times a word i was found in both the t and an h of a single entailment pair, while C_t and C_h were equal to the number of times a word was found in any t or h, respectively. A hand-crafted corpus of 100,000 entailment pairs was used to compute values for C_t, C_h, and C_{th}.

Features used in our classifier were selected from a number of sources, including (Hickl et al., 2006; Zanzotto et al., 2006; Glickman and Dagan, 2005).

A partial list of the features used in the Entailment Classifier used in our system is provided in Figure 3.

ALIGNMENT FEATURES: Derived from the results of the alignment of each pair of commitments performed during Commitment Selection.
◇1◇ LONGEST COMMON STRING: This feature represents the longest contiguous string common to both texts.
◇2◇ UNALIGNED CHUNK: This feature represents the number of chunks in one text that are not aligned with a chunk from the other
◇3◇ LEXICAL ENTAILMENT PROBABILITY: Defined as in (Glickman and Dagan, 2005).

DEPENDENCY FEATURES: Computed from the semantic dependencies identified by the PropBank- and NomBank-based semantic parsers.
◇1◇ ENTITY-ARG MATCH: This is a boolean feature which fires when aligned entities were assigned the same argument role label.
◇2◇ ENTITY-NEAR-ARG MATCH: This feature is collapsing the arguments Arg$_1$ and Arg$_2$ (as well as the Arg$_M$ subtypes) into single categories for the purpose of counting matches.
◇3◇ PREDICATE-ARG MATCH: This boolean feature is flagged when at least two aligned arguments have the same role.
◇4◇ PREDICATE-NEAR-ARG MATCH: This feature is collapsing the arguments Arg$_1$ and Arg$_2$ (as well as the Arg$_M$ subtypes) into single categories for the purpose of counting matches.

SEMANTIC/PRAGMATIC FEATURES: Extracted during preprocessing.
◇1◇ NAMED ENTITY CLASS: This feature has a different value for each of the 150 named entity classes.
◇2◇ TEMPORAL NORMALIZATION: This boolean feature is flagged when the temporal expressions are normalized to the same ISO 8601 equivalents.
◇3◇ MODALITY MARKER: This boolean feature is flagged when the two texts use the same modal verbs.
◇4◇ SPEECH-ACT: This boolean feature is flagged when the lexicons indicate the same speech act in both texts.
◇5◇ FACTIVITY MARKER: This boolean feature is flagged when the factivity markers indicate either TRUE or FALSE in both texts simultaneously.
◇6◇ BELIEF MARKER: This boolean feature is set when the belief markers indicate either TRUE or FALSE in both texts simultaneously.

Figure 3: Features used in the Entailment Classifier

6 Experiments and Results

We evaluated the performance of our commitment-based system for RTE against the 1600 examples found in the PASCAL RTE-2 and RTE-3 datasets.[7] Table 1 presents results from our system when trained on the 1600 examples taken from the RTE-2 and RTE-3 Test Sets.

Task	Length	IE	IR	QA	SUM	Total
RTE-2	Short	0.753	0.883	0.863	0.855	0.8385
RTE-3	Short	0.784	0.911	0.909	0.869	0.8331
RTE-3	Long	0.789	0.778	0.943	0.778	0.8290
Total	–	0.7690	0.8790	0.8890	0.8600	0.8493

Table 1: Performance of Commitment-based RTE.

Accuracy varied significantly (p <0.05) across each of the four tasks. Performance (in terms of accuracy and average precision) was highest on the QA set (88.9% accuracy) and lowest on the IE set (76.9%). The length of the *text* (either "short" or "long") did not significantly impact performance, however; in fact, as can be seen in Table 1, average accuracy was nearly the same for examples featuring "short" or "long" *texts*.

In order to quantify the impact that additional sources of training data could have on the performance an RTE system (and to facilitate comparisons with top systems like (Hickl et al., 2006), which were trained on tens of thousands of entailment pairs), we used the techniques described in (Bensley and Hickl, 2008) to generate a large training set of 100,000 text-hypothesis pairs in order to train our entailment classifier. Table 2 summarizes the performance of our RTE system on the RTE-2 and RTE-3 Test Sets when trained on increasing amounts of training data.[8]

Training Corpus	Accuracy	Average Precision
800 pairs (RTE-2 Dev)	0.8493	0.8611
10,000 pairs	0.8550	0.8742
25,000 pairs	0.8489	0.8322
50,000 pairs	0.8575	0.8505
100,000 pairs	0.8850	0.8785

Table 2: Impact of Training Corpus Size.

Unlike (Hickl et al., 2006), we experienced only a small increase (3%) in overall accuracy when training on increasingly larger corpora of examples. While large training corpora may provide an important source of knowledge for RTE, these results suggest that our commitment extraction-based approach may nullify the gains in performance seen by pure classification-based approaches. We believe that by training an entailment classification model based on the output of a commitment extraction module, we can reduce the number of deleterious features included in the model – and thereby, reduce the overall number of training examples needed to achieve the same level of performance.

In a second experiment, we investigated the performance gains that could be attributed to the choice of weighting function used to select commitments from a commitment set. In order to

[7] Data created for the PASCAL RTE-2 and RTE-3 challenges was organized into four datasets which sought to approximate the kinds of inference required by four different NLP applications: information extraction (IE), information retrieval (IR), question-answering (QA), and summarization (SUM). The RTE-3 Test Sets includes 683 "short" examples and 117 "long" examples; the RTE-2 Test Set includes 800 "short" examples.

[8] In the RTE evaluations, *accuracy* is defined as the percentage of entailment judgments correctly identified by the system. Average precision is defined as "the average of the system's precision values at all points in the ranked list in which recall increases".

Approach	Without Paraphrasing	With Paraphrasing	Δ
Term overlap (Zanzotto et al., 2006)	0.5950	0.6750	+0.0800
Approximate Tree Edit Distance (Schilder and McInnes, 2006)	0.6550	0.5933	-0.0617
LEP (Glickman et al., 2005)	0.6000	0.6800	+0.0800
Lexical Similarity (Adams, 2006)	0.6200	0.6788	+0.0588
Graph Matching (MacCartney et al., 2006)	0.6433	0.6533	+0.0100
Classification-Based Alignment (Hickl et al., 2006)	0.7650	0.7700	+ 0.0050
Structured Prediction-Based Alignment (Taskar et al., 2005a)	0.7900	0.8493	+ 0.0593

Table 3: Impact of Commitment Selection Learning.

perform this comparison, we implemented a total of 7 different functions previously investigated by teams participating in the previous PASCAL RTE Challenges, including (1) a simple term-overlap measure introduced as a baseline in (Zanzotto et al., 2006), (2) the approximate tree edit distance metric used by (Schilder and McInnes, 2006), (3) (Glickman et al., 2005)'s measure of lexical entailment probability (LEP), (4) the lexical similarity measure described in (Adams, 2006), (5) our interpretation of the semantic graph-matching approach described in (MacCartney et al., 2006), (6) the classification-based term alignment approach described in (Hickl et al., 2006), and (7) the structured prediction-based alignment approach introduced in this paper. Results from this 7-way comparison are presented in Table 3.

While we found that the choice of mechanism used to weight commitments did significantly impact RTE performance ($p < 0.05$), the inclusion of generated paraphrases appeared to only have a slight positive impact on overall performance, boosting RTE accuracy by an average of 3.1% (across the 7 methods), and by a total of 5.9% in the approach we describe in this paper. While paraphrasing can be used to enhance the performance of some current RTE systems (see performance of the Tree Edit Distance-based RTE system for a case where paraphrasing negatively impacted performance), realized gains are still relatively modest across most approaches.

In a third experiment, we found that RTE performance depended on both the type – and the number – of commitments extracted from a *t-h* pair, regardless of the learning algorithm used in commitment selection. Table 4 presents results from experiments when (1) no commitment extraction was conducted, (2) extraction strategies were run in isolation[9], (3) combinations of extraction strategies were considered, or (4) all of possible extraction strategies (listed in Section 3) were considered. The best-performing condition for each RTE strategy is presented in **bold**.

Increasing the amount of linguistic knowledge available from a commitment set did significantly ($p < 0.05$) impact the performance of RTE, regardless of the actual learning algorithm used to compute the likelihood of an entailment relationship. Combining all four extraction strategies proved best for four RTE approaches (Term Overlap, Lexical Entailment Probability, Graph Matching, and Structured Prediction-based Alignment). Including coreference-based commitments reduced the accuracy of two RTE strategies (Lexical Similarity and Classification-based Alignment). This is most likely due to the fact that coreference-based commitments reincorporate the antecedent of pronouns and other referring expressions into the generated commitments, thereby adding additional lexical information to the sets of features used in computing entailment.

7 Conclusions

This paper introduced a new framework for recognizing textual entailment which depends on the extraction of the discourse commitments that can be inferred from a conventional interpretation of a text passage. By explicitly enumerating the set of inferences that can be drawn from a t or h, our approach is able to reduce the task of RTE to the identification of the set of commitments that support the inference of each corresponding commitment extracted from a hypothesis. This approach correctly classified more than 80% of examples from the PASCAL RTE Test Sets, without the need for additional sources of training data.

Acknowledgments

This material is based upon work funded in whole or in part by the U.S. Government and any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government.

References

Adams, Rod. 2006. Textual entailment through extended lexical overlap. In *Proceedings of the Second PASCAL Recognising Textual Entailment Challenge (RTE-2)*.

[9] Four strategies were considered: (1) syntactic decomposition (*SD*), (2) supplemental expression (*SE*) extraction, (3) coreference resolution (*Coref*), (4) and paraphrasing (*Para*).

	None	Individual Strategies				Strategy Combinations			All
		SD	SE	Coref	Para	SD+SE	SD,SE,Coref	SD,SE,Para	
Term Overlap	0.5708	0.5750	0.5850	0.5758	0.5925	0.5950	0.5925	0.6500	**0.6750**
Approximate Tree Edit Distance	0.5117	0.6158	0.6175	0.5142	0.5158	**0.6592**	0.6583	0.6108	0.5930
Lexical Entailment Probability	0.6067	0.6300	0.5950	0.5942	0.6475	0.6292	0.6000	0.6658	**0.6800**
Lexical Similarity	0.5833	0.5875	0.5817	0.5850	0.5975	0.6100	0.6200	**0.6825**	0.6800
Graph Matching	0.5850	0.6667	0.6450	0.6483	0.6017	0.6317	0.6430	0.6208	**0.6530**
Classification-Based Alignment	0.6692	0.6825	0.6700	0.6758	0.6842	0.7508	0.7492	**0.7992**	0.7700
Structured Prediction-Based Alignment	0.6750	0.7083	0.6875	0.6675	0.6742	0.7492	0.7900	0.7842	**0.8493**

Table 4: Impact of Commitment Availability.

Bensley, Jeremy and Andrew Hickl. 2008. Unsupervised Resource Creation for Textual Inference Applications. In *LREC 2008*, Marrakech.

Bos, Johan and Katya Markert. 2006. When logical inference helps in determining textual entailment (and when it doesn't). In *Proceedings of the Second PASCAL Recognizing Textual Entailment Conference*, Venice, Italy.

Curran, James, Stephen Clark, and Johan Bos. 2007. Linguistically Motivated Large-Scale NLP with C-and-C and Boxer. In *ACL 2007 (Demonstration Session)*, Prague.

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

Dice, L.R. 1945. Measures of the Amount of Ecologic Association Between Species. In *Journal of Ecology*, volume 26, pages 297–302.

Glickman, Oren and Ido Dagan. 2005. A Probabilistic Setting and Lexical Co-occurrence Model for Textual Entailment. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, Ann Arbor, USA.

Glickman, Oren, Ido Dagan, and Moshe Koppel. 2005. Web based textual entailment. In *Proceedings of the First PASCAL Recognizing Textual Entailment Work-shop*.

Gunlogson, Christine. 2001. *True to Form: Rising and Falling Declaratives as Questions in English*. Ph.D. thesis, University of California, Santa Cruz.

Haghighi, Aria, 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.

Hickl, Andrew, 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*.

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

Jijkoun, V. and M. de Rijke. 2005. Recognizing Textual Entailment Using Lexical Similarity. In *Proceedings of the First PASCAL Challenges Workshop*.

MacCartney, Bill, 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.

Mushegian, Arcady and Eugene Koonin. 1996. 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.

Pedersen, T., S. Patwardhan, and J. Michelizzi. 2004. WordNet::Similarity - Measuring the Relatedness of Concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, San Jose, CA.

Potts, Christopher, editor, 2005. *The Logic of Conventional Implicatures*. Oxford University Press.

Schilder, F. and B. Thomson McInnes. 2006. TLR at DUC 2006: Approximate Tree Similarity and a New Evaluation Regime. In *Proceedings of HLT-NAACL Document Understanding Workshop (DUC 2006)*.

Stalnaker, Robert, 1979. *Assertion*, volume 9, pages 315–332.

Szpektor, Idan, Eyal Shnarch, and Ido Dagan. 2007. Instance-based evaluation of entailment rule acquisition. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 456–463, Prague, Czech Republic, June. Association for Computational Linguistics.

Taskar, Ben, Simone Lacoste-Julien, and Michael Jordan. 2005a. Structured prediction via the extragradient method. In *Proceedings of Neural Information Processing Systems*.

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

Vanderwende, Lucy, Arul Menezes, and Rion Snow. 2006. Microsoft Research at RTE-2: Syntactic Contributions in the Entailment Task: an implementation. In *Proceedings of the Second PASCAL Challenges Workshop*.

Wong, Yuk Wah and Raymond Mooney. 2007. Learning synchronous grammars for semantic parsing with lambda calculus. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 960–967, Prague, Czech Republic, June. Association for Computational Linguistics.

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

Zettlemoyer, L. S. and M. Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of UAI-05*.