A Discourse Commitment-Based Framework for Recognizing Textual Entailment

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

Abstract

In this paper, we introduce a new framework for recognizing textual entailment 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 or a hypothesis. 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*. Promising results were achieved: our system correctly identified more than 80% of examples from the RTE-3 Test Set correctly, without the need for additional sources of training data or other web-based resources.

1 Introduction

Systems participating in the previous two PASCAL Recognizing Textual Entailment (RTE) Challenges (Bar-Haim et al., 2006) 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), (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) or model building systems (Bos and Markert, 2006).

While 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. 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 set of propositions which are inferable – whether asserted, presupposed, or conventionally or conversationally implicated – from a text-hypothesis pair.

In this paper, we introduce a new framework for recognizing textual entailment which depends on extraction of the set of publicly-held beliefs – or *discourse commitments* – that can be ascribed to the author of a text or a hypothesis. We show that once a set of discourse commitments have been extracted from a text-hypothesis pair, the task of recognizing textual entailment can be reduced to the identification of the one (or more) commitments from the text which are most likely to support the inference of each commitment extracted from the hypothesis. More 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.

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-3 Challenge. Sections 3, 4, and 5 describe details of our systems for Commitment Extraction, Commitment Se-

Proceedings of the Workshop on Textual Entailment and Paraphrasing, pages 171–176,
Prague, June 2007.

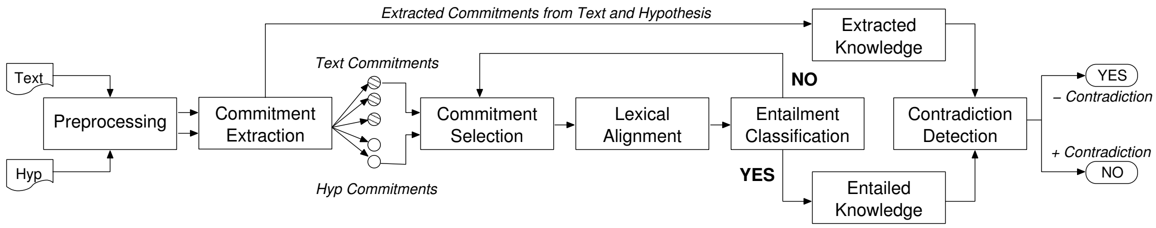

Figure 1: System Architecture.

lection, and Entailment Classification, respectively. Finally, Section 6 discusses results from this year's evaluation, and Section 7 provides our conclusions.

2 System Overview

The architecture of our system for recognizing textual entailment (RTE) is presented in Figure 1.

In our system, *text-hypothesis* (*t-h*) pairs are initially submitted to a *Preprocessing* module which (1) syntactic parses each passage (using an implementation of the (Collins, 1999) parser), (2) identifies semantic dependencies (using a semantic dependency parser trained on PropBank (Palmer et al., 2005) and NomBank (Meyers et al., 2004)), (3) annotates named entities (using LCC's *CiceroLite* named entity recognition system), (4) resolves instances of pronominal and nominal coreference (using a system based on (Luo et al., 2004)), and (5) normalizes temporal and spatial expressions to fully-resolved instances (using a technique first introduced in (Aarseth et al., 2006)).

Annotated passages are then sent to a *Commitment Extraction* module, which uses a series of extraction heuristics in order to enumerate a subset of the discourse commitments that are inferable from either the *text* or *hypothesis*. Following (Gunlogson, 2001; Stalnaker, 1979), we assume that a discourse commitment (c) represents the any of the set of propositions that can necessarily be inferred to be true, given a conventional reading of a text passage. The complete list of commitments that our system is able to extract from from the t used in examples 34 and 36 from the RTE-3 Test Set is presented in Figure 2. (Details of our commitment extraction approach are presented in Section 3.)

Commitments are then sent to a *Commitment Selection* module, which uses a weighted bipartite matching algorithm first described in (Taskar et al., 2005b) in order to identify the commitment from the t which features the best alignment for each commitment extracted from the h. The commitment pairs identified for the hypotheses from 34 and 36 are highlighted in Figure 2. (Details of our method for selecting and aligning commitments are provided in Section 4.)

Each pair of commitments are then considered in turn by an *Entailment Classification* module, which follows (Bos and Markert, 2006; Hickl et al., 2006) in using a decision tree classifier in order to compute the likelihood that a commitment extracted from a t textually entails a commitment extracted from an h.

If a commitment pair is judged to be a positive instance of TE, it is sent to an *Entailment Validation* module, which uses a system for recognizing instances of textual contradiction (RTC) based on (Harabagiu et al., 2006) in order to determine whether the (presumably) entailed hypothesis is contradicted by any of other commitments extracted from the t during commitment extraction. If no text commitment can be identified which contradicts the hypothesis, it is presumed to be textually entailed, and a judgment of YES is returned. Alternatively, if the entailed h is textually contradicted by one (or more) of the commitments extracted from the t, the h is considered to be contradicted by the t, the entailment pair is classified as a negative instance of TE, and a judgment of NO is returned.

In contrast, when commitment pairs are judged to be negative instances of TE by the Entailment Classifier, the current pair is removed from further consideration by the system, and the next most likely commitment pair is considered. Commitment pairs are considered in decreasing order of the probability output by the Commitment Selection module until a positive instance of TE is identified – or until there are no more commitment pairs with a selection probability greater than a pre-defined threshold.

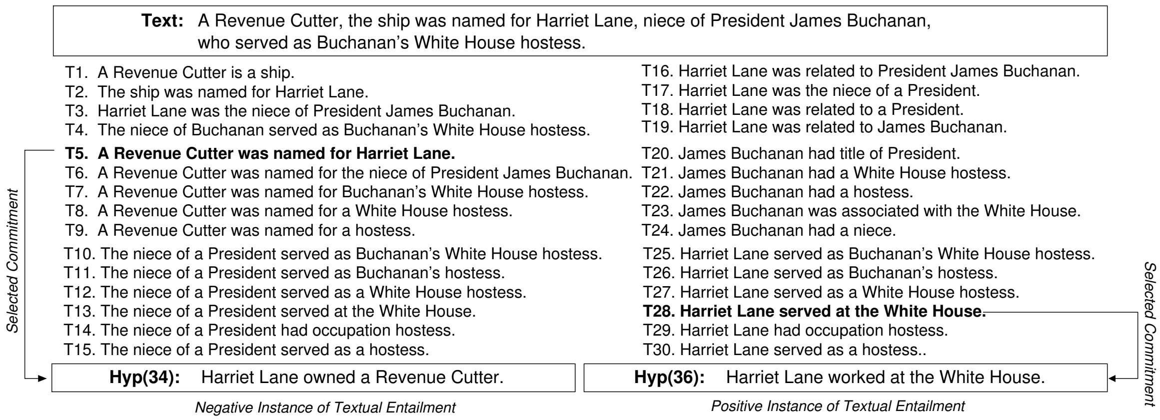

Figure 2: Text Commitments Extracted from Examples 34 and 36.

3 Extracting Discourse Commitments

Following Preprocessing, our system for RTE leverages a series of heuristics in order to extract a subset of the discourse commitments available from a text-hypothesis pair. In this section, we outline the five classes of heuristics we used to extract commitments for the RTE-3 Challenge.

Sentence Segmentation: We use a sentence segmenter to break text passages into sets of individual sentences; commitments are then extracted from each sentence independently.

Syntactic Decomposition: We use heuristics to syntactically decompose sentences featuring coordination and lists into well-formed sentences that only include a single conjunct or list element.

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. In our current system, heuristics were used to extract supplemental expressions from each sentence under consideration and to create new sentences which specify the CI conveyed by the expression.

Relation Extraction: We used an in-house relation extraction system to recognize six types 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. Again, as with supplemental expressions, heuristics were used to generate new commitments which expressed the semantics conveyed by these nominal relations.

Coreference Resolution: We used systems for resolving 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. *Harriet Lane*, *the niece*, *Buchanan's White House hostess*), new commitments were generated from the existing set of commitments which incorporated each co-referential mention.

4 Commitment Selection

Following Commitment Extraction, we used an word 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 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.[1] As with (Taskar et al., 2005b), we use the large-margin structured prediction model

[1] 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*.

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 (1).

$$y_i = \arg\max_{\bar{y}_i \in Y} w^\top f(x_i, \bar{y}_i), \forall i \quad (1)$$

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 named entity similarity equality), and (3) word association features (computed using the Dice coefficient (Dice, 1945)[2]). In order to provide a training set which most closely resembled the RTE-3 Test Set, we hand-annotated token alignments for each of the 800 entailment pairs included in the 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, 2005) 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.

We compared the performance of our word alignment and commitment selection algorithms against an implementation of the lexical alignment classifier described in (Hickl et al., 2006) on commitments extracted from the entailment pairs from the RTE-2 Test Set. Table 1 presents results from evaluations of these two models on the token alignment and commitment selection tasks. (Gold standard annotations for each task were created by hand by a team of 3 annotators following the RTE-3 evaluations.)

[2]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}.

Task	Measurement	Current Work	Hickl et al.
Token Alignment	Precision	94.55%	92.22%
Token Alignment	MRR	0.9219	0.8797
Commitment Selection	Precision	89.50%	72.50%
Commitment Selection	MRR	0.8853	0.7410

Table 1: Alignment and Selection Performance

5 Entailment Classification

Following work done by (Bos and Markert, 2006; Hickl et al., 2006) for the RTE-2 Challenge, we used a decision tree (C5.0 (Quinlan, 1998)) to estimate the likelihood that a commitment pair represented a valid instance of textual entailment.[3] Confidence values associated with each leaf node (i.e. YES or NO) were normalized and used to rank examples for the official submission.

In a departure from previous work (such as (Hickl et al., 2006)) which leveraged large corpora of entailment pairs to train an entailment classifier, our model was only trained on the 800 text-hypothesis pairs found in the RTE-3 Development Set (DevSet). Features were selected manually by performing ten-fold cross validation on the DevSet. Maximum performance of the entailment classifier on the DevSet is provided in Table 2.

	IE	IR	QA	SUM	Total
Accuracy	0.8450	0.8750	0.8850	0.8600	0.8663
Average Precision	0.8522	0.8953	0.9005	0.8959	0.8860

Table 2: Entailment Classifier Performance.

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

6 Experiments and Results

We submitted one ranked run in our official submission for this year's evaluation. Official results from the RTE-3 Test Set are presented in Table 3.

	IE	IR	QA	SUM	Total
Accuracy	0.6750	0.8000	0.9000	0.8400	0.8038
Average Precision	0.7760	0.8133	0.9308	0.8974	0.8815

Table 3: Official RTE-3 Results.

Accuracy and average precision 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 (90.0% precision) and lowest on the IE set (67.5%).

The length of the *text* (either *short* or *long*) did not significantly impact performance, however; in fact,

[3]We used a pruning confidence of 20% in our model.

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 9000 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

as can be seen in Table 4, total accuracy was nearly the same for examples featuring *short* or *long texts*.

	Short		Long	
	n	Accuracy	*n*	Accuracy
IE	181	0.6685	19	0.7368
IR	146	0.8082	54	0.7778
QA	165	0.8909	35	0.9429
SUM	191	0.8482	9	0.6667
Total	683	0.8023	117	0.8120

Table 4: Short vs. Long Pairs.

In experiments conducted following the RTE-3 submission deadline, we found that using a system for recognizing textual contradiction to validate judgments output by the entailment classifier had only a slight positive impact on the overall performance of our system. Table 5 compares performance of our RTE system when four different configurations of our system for recognizing textual contradiction was used.

When used with its default threshold (λ = 0.85), we discovered that using textual contradiction enabled us to identify 17 additional examples (2.13% overall) that were not available when using our system for RTE alone.[4] When we hand-tuned λ to maximize performance on the RTE-3 Test Set, we found that accuracy could be increased by 3.0% over the baseline (to 81.25% overall). Despite its limited effectiveness on this year's Test Set, we believe that net positive effect of using textual contradiction to validate textual entailment judgments suggests that this technique has merit and should be explored in future evaluations.

Validation?	λ	IE	IR	QA	SUM	Total
Yes *(RTE-3)*	0.85	0.6750	0.8000	0.9000	0.8400	0.8038
Yes	0.75	0.6900	0.8100	0.8850	0.8650	0.8125
Yes	0.65	0.6550	0.8000	0.8850	0.8250	0.7913
No	–	0.6550	0.8000	0.8650	0.8250	0.7865

Table 5: Impact of Validation.

In a second post hoc experiment, we sought to quantify the impact that additional sources of training data could have on the performance of our RTE system. Although our official submission was only trained on the 800 *t-h* pairs found in the RTE-3 Development Set, we followed (Hickl et al., 2006) in using a large, hand-crafted training set of 100,000 text-hypothesis pairs in order to train our entailment classifier. Even though previous work has shown that RTE accuracy increased with the size of the training set, our experiments showed no correlation between the size of the training corpus and the overall accuracy of the system. Table 6 summarizes the performance of our RTE system when trained on increasing amounts of training data. While increasing the training data to approximately 10,000 training examples did positively impact performance, we discovered that using a training corpus of a size equal to (Hickl et al., 2006)'s had nearly no measurable impact on the observed performance of our system.

Training Corpus	Accuracy	Average Precision
800 pairs (RTE-3 Dev)	0.8038	0.8815
10,000 pairs	0.8150	0.8939
25,000 pairs	0.8225	0.8834
50,000 pairs	0.8125	0.8355
100,000 pairs	0.8050	0.8003

Table 6: Impact of Training Corpus Size.

While large training corpora (like (Hickl et al., 2006)'s or the one compiled for this work) may provide an important source of lexico-semantic information that can be leveraged in performing an entailment classification, these results suggest that our approach based on commitment extraction may nullify

[4] We learned the default threshold by training on the textual contradiction corpus compiled by (Harabagiu et al., 2006).

the gains in performance seen by these approaches.

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. In our current work, we show that this approach can be used to correctly classify more than 80% of examples from the RTE-3 Test Set, without the need for additional sources of training data or web-based resources.

References

Paul Aarseth, John Lehmann, Murat Deligonul, and Luke Nezda. 2006. TASER: A Temporal and Spatial Expression Recognition and Normalization System. In *Proceedings of the Automatic Content Extraction (ACE) Conference*.

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*.

Johan Bos 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.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, Univ. of Pennsylvania.

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

Oren Glickman 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.

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

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, 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*.

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

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

X. Luo, A. Ittycheriah, H. Jing, N. Kambhatla, and S. Roukos. 2004. A mention-synchronous coreference resolution algorithm based on the Bell Tree. In *Proceedings of the ACL-2004*, Barcelona, Spain.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The nombank project: An interim report. In A. Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA, May 2 - May 7. 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.

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

T. Pedersen, 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.

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

R. Quinlan. 1998. C5.0: An Informal Tutorial. RuleQuest.

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

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

Ben Taskar, 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)*, Vancouver, Canada.

Lucy Vanderwende, 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*.