Deciding Entailment and Contradiction with Stochastic and Edit Distance-based Alignment

Sebastian Padó, Marie-Catherine de Marneffe, Bill MacCartney,
Anna N. Rafferty, Eric Yeh, and Christopher D. Manning

Stanford University

January 15, 2009

This paper describes the Stanford submission to the TAC 2008 RTE track.

1 The Stanford RTE system

Our experiments are based on the Stanford textual entailment recognition system (MacCartney et al., 2006). The input to the system is a passage/hypothesis pair. The system uses a three-stage architecture that (a) conducts linguistic analysis, (b) builds an alignment between dependency graphs of the two segments, and (c) performs inference to determine entailment.

Analysis stage. Our goal at this stage is to compute linguistic representations of the passage and the hypothesis that contain as much information as possible about their semantic content. We use typed dependency graphs generated by the Stanford parser (Klein and Manning, 2003; de Marneffe et al., 2006), which contain a node for each word and labeled edges representing the grammatical relations between words. Named entities are identified by a CRF-based NER system, and contiguous collocations which appear in WordNet are collapsed.

Alignment stage. The second stage identifies a good partial alignment between the graphs representing the hypothesis and the passage. There are many ways in which such an alignment can be found. The Stanford system constructs the alignment as the highest-scoring mapping from each node in the hypothesis graph to a single node in the passage graph, or to null. We choose a locally decomposable scoring function, such that the score of an individual alignment $s(p, h, a)$ is the sum of the local node score (s_w) and edge alignment score (s_e):

$$s(p,h,a) = \sum_{i \in h} s_w(h_i, a(i)) + \sum_{(i,j) \in e(h)} s_e((h_i, h_j), (a(h_i), a(h_j)))$$

where we use the notation $a(x)$ to refer to the word in the passage which is aligned to the word x in the hypothesis under the alignment a, and $e(x)$ to refer to a function returning the set of edges in a hypothesis x. The first term is the sum of the alignment scores of the individual words, which combine lexical similarity judgments from about ten resources, including WordNet, InfoMap, Dekang Lin's thesaurus, and gazetteers. The second term is the sum of the alignment scores of the pairs of words which are connected by an edge in the hypothesis graph.

The large number of possible alignments (exponential in the number of hypothesis words) makes exhaustive search intractable. Instead, we use a stochastic search technique based on Gibbs sampling, a well-known Markov Chain Monte Carlo technique (see de Marneffe et al. (2007) for details). This Gibbs sampler is guaranteed to give us samples from the posterior distribution over alignments as defined by the scoring function.

Inference stage. The final stage determines if the hypothesis is entailed by the passage. We construct a set of features based on the output of the previous stages. These features model a wide range of syntactic, lexical, and semantic phenomena, including factivity, polarity, antonymy, felicity of adjuncts, modality, quantification, matches and mismatches of names as well as of dates and numbers, compatibility of syntactic structure, and the quality of the alignment. The final score for each passage-hypothesis pair is computed as the dot product of the feature values and a weight vector. The feature weights, as well as the decision boundary between entailment and non-entailment, can either be set by hand, or learned from a development dataset with (regularized) logistic regression.

The disassociation between an alignment and an entailment stage is in contrast to most systems developed for RTE, which determine entailment directly from (graph) alignment quality. We have found that alignment scoring and deciding entailment are two conceptually distinct tasks, whose separation is important to deal with a broad class hypotheses that align well with passages but are still not entailed by them. Consider the following example:

[RTE2-dev 156]

P: Energy analysts said oil prices could soar as high as $80 a barrel and drivers in the U.S. could soon be paying $3 a gallon for gasoline, if damage reports from oil companies bear bad news.

H: Oil prices surged.

In this passage/hypothesis pair, the hypothesis aligns perfectly with the passage, but its embedding in a conditional on the passage side cancels out the entailment (see MacCartney et al. (2006) for details and more examples). Contradictions are another important class where a good alignment quality does not equate entailment (see Section 3 below).

Results. Table 1 contains the results for this year's Stanford submissions on the RTE-4 test set. The results for the system as described above are shown in the first row (Stanford 1 - Stochastic aligner). The 2-way accuracy is somewhat higher than for the most similar system Stanford submitted to last year's RTE-3 competition (Stanford Core + coref) which obtained 60.5% accuracy on the test set. However, average precision is low at 44%.

Our preliminary analysis of the RTE-4 results has identified the different properties of the RTE1-3 datasets and the RTE-4 dataset as a major problem for our system. In particular, the relation between the length of the passage and hypothesis has changed significantly, from

Submission	Alignment	2-way Acc.	3-way Acc.	Average Precision
Stanford 1	Stochastic	61.4	55.3	44.2
Stanford 2	MANLI	57.0	50.1	54.3

Table 1: Results for two submitted runs on the RTE-4 test set (Stochastic vs. MANLI aligner)

2:1 (RTE-1) or 3:1 (RTE-3) to above 4:1 this year. As a result, it was "too easy" for our system to align hypothesis material to the passage: the linear classifier in the inference stage, whose weights were optimized on RTE1-3, initially classified more than 65% of the RTE-4 instances as entailments. This problem also results in faulty confidence estimates which lead to low average precision numbers.

2 Research Focus 1: Comparing alignment strategies

2.1 Motivation and Description

One of our research goals for RTE-4 was to experiment with integrating a new phrase-based alignment system known as MANLI (MacCartney et al., 2008) to replace the Stochastic aligner described in Section 1. The MANLI aligner was first developed independently of the RTE system as a standalone alignment system. It uses supervised training on gold-standard alignment annotations generated by Microsoft Research (MSR) for the RTE-2 data (Brockett, 2007), and it was designed to optimize accuracy in recovering those gold alignments, rather than to optimize the accuracy of a complete RTE system on the primary RTE task of predicting inferential validity. Stochastic aligner in evaluations on the MSR data, we were hopeful that it could also lead to better results on the primary RTE task. In particular, the Stochastic aligner is limited to constructing one-to-one token alignments, which impedes its ability to match multi-word phrases; the MANLI aligner aims to remove this barrier. The MANLI system consists of four main elements:

A phrase-based alignment representation. MANLI represents an alignment by a set of edits (substitutions, deletions, and insertions) whose arguments are phrases (contiguous token spans). Using a phrase-based representation permits us to align phrases (such as *started crying* and *burst into tears*) as units, without being forced to make an arbitrary choice as to which word goes with which word. Moreover, our scoring function can make use of lexical resources which have information about semantic relatedness of multi-word phrases, not merely individual words.

A feature-based scoring function. MANLI uses a simple feature-based linear scoring function, in which the score of an alignment is the sum of the scores of the edits it contains, and the score of an edit is the dot product of a vector encoding its features and a vector of weights. The features for each edit encode its type and size, and whether it involves non-constituents. For substitution edits, the features also include a lexical similarity score computed as a max over a variety of component similarity functions, many based on external lexical resources, including both manually compiled resources (such as WordNet and NomBank) and automatically induced resources (such as Dekang Lin's distributional similarity

		Stochastic	
		right	wrong
MANLI	right	460	110
	wrong	154	276

Table 2: Stochastic Aligner vs. MANLI Aligner on two-way task

scores). Substitution edits also use contextual features, including a distortion score and a matching-neighbors feature.

Decoding using simulated annealing. For decoding, MANLI uses a simulated annealing strategy. Beginning from an arbitrary alignment, we make a series of local steps, at each iteration sampling from a set of possible successors according to scores assigned by our scoring function. The sampling is controlled by a "temperature" which falls over time. At the beginning of the process, successors are sampled with nearly uniform probability, which helps to ensure that the space of possibilities is explored and local maxima are avoided. As the temperature falls, there is a ever-stronger bias toward high-scoring successors, so that the algorithm converges on a near-optimal alignment.

Averaged perceptron learning. To tune the model parameters, we use an adaptation of the averaged perceptron algorithm (Collins, 2002). After initializing the weights to 0, we perform a fixed number of training epochs. In each epoch, we iterate through the training data, updating the weight vector at each training example according to the difference between the features of the target alignment and the features of the alignment produced by the decoder using the current weight vector. The size of the update is controlled by a learning rate which decreases over time. At the end of each epoch, the weight vector is normalized and stored. The final result is the average of the stored weight vectors, omitting vectors from a fixed number of epochs at the beginning of the run, which tend to be of poor quality.

2.2 Results and Analysis

The output of the Stanford system using the MANLI aligner was submitted as run Stanford-2 (cf. Table 1). The results did not completely meet our expectations, but are overall promising for MANLI. In terms of accuracy, MANLI still trails the Stochastic aligner by around 4.5%. We think that this is due to the RTE system features being optimized for the Stochastic aligner, instead of the substantially different edit-sequence based alignment, and the fact that MANLI currently only has access to very limited amounts of paraphrase information. On the other hand, MANLI outperforms the Stochastic aligner on average precision by 10%. It appears that edit sequence-based scoring is able to better rank instances by confidence.

Table 2 compares the performance of the two alignment models on the RTE-4 dataset. Note that their answers differ for around 26% of datapoints. In general, the MANLI systems tends to be more cautious in aligning non-literal correspondences than the Stochastic aligner. Figure 1 shows two example sentences that illustrate the difference the between the two systems. In the first sentence, 634, the Stochastic aligner identifies the correspondence between the main predicates (*underscore/highlight*) and aligns them, contributing to the

	Sentence pair	Stochastic	MANLI
634 P:	The moderate earthquake that struck the LA area yesterday **underscores** the importance of preparation to prevent damage, death and injuries in the event of a bigger temblor, officials said.	ENTAIL correct	UNKNOWN incorrect
H:	A moderate quake in LA **highlights** the need for preparation.		
194 P:	Japan has **moved** to impose additional economic sanctions against North Korea in response to its claimed nuclear test earlier this week. [...]	ENTAIL incorrect	UNKNOWN correct
H:	Japan **lifts** sanctions against North Korea.		

Figure 1: RTE-4 sentence pairs differing in answers from the Stochastic and MANLI aligners

(correct) prediction of entailment. This link is not found by the MANLI aligner, which leads to a low score due to a missing alignment between the sentence roots. The second sentence shows a similar situation. Here, however, the alignment that the Stochastic aligner constructs between *move* to *lift* leads it onto the wrong track to predict entailment where the right answer in the two-way track would have been no entailment. This is correctly recognized by the MANLI aligner.

3 Research Focus 2: Contradiction detection

3.1 Motivation and Description

In the 3-way task, the system also needs to decide whether pairs of texts are contradictory. To do so, we use the Stanford contradiction detection system (de Marneffe et al., 2008), which is an adaptation of our RTE system. The decoupling of alignment score and entailment decision allows us to integrate both entailment and contradiction detection within the same architecture: linguistic analysis and alignment are shared by both systems, but feature extraction differs. In the final stage of the contradiction detection system, we extract contradiction features on which we apply logistic regression to classify the pair as contradictory or not. Features weights are hand-set, guided by linguistic intuition.

Filtering non-coreferent events. Contradiction features rely on mismatches between the passage and the hypothesis. However pairs of sentences which do not describe the same event, and thus cannot be contradictory to one another, could nonetheless contain mismatching information. An extra stage to filter non-coreferent events is therefore added before feature extraction. For example, in the following pair, it is necessary to recognize that *the Johnstown Flood* has nothing to do with *a ferry sinking*; otherwise conflicting death tolls (*2,000* vs. *100 or more*) result in labeling the pair a contradiction.

P: *More than 2,000 people* lost their lives in the devastating Johnstown Flood.

H: *100 or more people* lost their lives in a ferry sinking.

	Precision	Recall
contradiction 3-ways	28.6	8.0
contradiction standalone	26.3	10.0

Table 3: Contradiction results of the contradiction detection system using the Stochastic alignment: integrated with entailment decision (3-ways) and contradiction detection alone.

This issue does not arise for the entailment/non-entailment decision: elements in the hypothesis that are not supported by the passage tend to lead to non-entailment irrespective of whether the same event is described. For contradiction, however, it is critical to focus attention on related sentences to avoid tagging sentences as contradictory that are merely about different events. We therefore discard all sentence pairs whose passage and hypothesis are not about the same event: this decision is based on topicality by looking at aligned NPs. The remaining pairs are then analyzed for specific evidence of contradiction.

Contradiction features. Mismatching information between sentences is often a good cue of non-entailment (Vanderwende et al., 2006), but it is not sufficient for contradiction detection which requires more precise comprehension of the consequences of sentences. Some of the features used in the Stanford RTE system have been more precisely defined to only capture mismatches in similar contexts, instead of global mismatching. These features, described in details in de Marneffe et al. (2008), include polarity differences, presence of antonyms, numeric mismatches, structural differences, modality, factivity and hand-coded contradictory relations (*Fernandez, of FEMA, . . .* is contradictory to *Fernandez doesn't work for FEMA*).

Combining the contradiction and entailment decisions. Since contradiction detection is not very reliable (as shown by results obtained on previous RTE datasets (de Marneffe et al., 2008) as well as on recent work on contradiction between functional relations (Ritter et al., 2008), we trust the entailment system more. After running both systems independently on the data, pairs are tagged as entailments if the entailment systems returns a positive verdict. Otherwise, pairs recognised as contradictions by the contradiction system are marked as such. All remaining sentences receive the label "unknown".

3.2 Results and Analysis

There were 150 contradictions in RTE-4. Table 3 shows results for the contradiction detection in the 3-way task and results for the standalone contradiction detection system (with Stochastic alignment). Recall is quite low. We see three reasons for this. First the event filter has to be improved: 47 "true" contradictions were filtered out by the event filter. This can be due to the difference in datasets as explained above, since the threshold for filtering was optimized on previous RTE corpora. Second, in the 3-way task, 3 "true" contradictions correctly identified by the contradiction detection system were discarded because they were tagged as entailments. It might be possible to devise a better combination of the two systems by including some high-precision features of the contradiction detection system into the entailment decision. Third, a little over half of the contradiction pairs in the dataset (53%)

require much deeper lexical knowledge than the system has, as these examples show:

[RTE4 42]

P: President Yar'Adua immediately announced an ambitious seven point plan, aimed at tackling the energy crisis, reducing unemployment, investing in agriculture and land reform, fighting crime as well as improving education and public transport. Nowadays, Nigeria enjoys a constant supply of electricity and power shortages *are a thing of the past*.

H: Nigeria power shortage is *to persist*.

[RTE4 51]

P: Jews migrated to Bahrain in the 19th century, mostly from Iran and Iraq. Their numbers increased early in the 20th century but decreased after the 1948 Arab-Israeli war, when many left for Israel, the U.S. and Europe. Jews keep a low profile in Bahrain, working mostly in banks, commercial and trade companies and retail. Jews are not allowed to work for the government or to *represent the nation*.

H: Bahrain names a Jewish *ambassador*.

[RTE4 64]

P: [. . .] No children were *among the victims*.

H: A French train crash *killed* children.

[RTE4 69]

P: [. . .] The report of a crash was a *false alarm*.

H: A plane crashes in Italy.

[RTE4 76]

P: [. . .] The current food crisis *was ignored*.

H: UN summit *targets* global food crisis.

[RTE4 95]

P: [. . .] The toilet of the International Space Station is *working perfectly*.

H: The space station has toilet *trouble*.

Achieving good precision on contradiction is tricky. Prominent causes of erroneously identified contradictions (30 errors in the 3-way task) are wrong structural mismatches (40%) as well as bad alignments between passage and hypothesis leading to features firing incorrectly (23%). In the following pair, the erroneous alignment of *member* in the hypothesis to *no member* in the passage leads to the recognition of a spurious polarity difference:
[RTE4 762]

P: At the present time, the Japanese imperial household has no male *member* under the age of 20; present law only allows male descendants to ascend to the Chrysanthemum Throne. The proposed solution would put Princess Aiko, the sole child of Crown Prince Naruhito and Crown Princess Masako in place as the most likely to inherit the position of monarch.

H: Princess Aiko is a *member* of the Japanese imperial household.

Other features that fire incorrectly are antonymy (10%), negation (10%), relations (6%) and numeric mismatch (3%). Errors in coreference resolution also lead to mistakes in contradiction detection (6%). The following pairs are examples of such errors:

[RTE4 899] (*Imports* and *exports* are identified as antonyms, but the verb argument structures are not taken into account)

P: The company affected by this ban, Flour Mills of Fiji, exports nearly US$900,000 worth of biscuits to Vanuatu yearly.

H: Vanuatu imports biscuits from Fiji.

[RTE4 391] (The verb *like* is interpreted as a polarity difference)

P: It is hard to *like* Will Carling. Honestly, that has nothing to do with the fact that he is captain of the England rugby team (again), but a lot to do with his status as an insufferably rude brat.

H: *Nobody likes* Will Carling, because he's rude.

[RTE4 332] (Numeric mismatch between *113* and *109*: the system cannot deal with arithmetical expressions like *109 people and four workers*)

P: The New York-bound Concorde crashed in a ball of fire shortly after takeoff from Paris Charles de Gaulle airport on July 25, 2000, killing all 109 people on board and four workers on the ground.

H: The crash killed 113 people.

4 Research Focus 3: Nonlocal information recovery

4.1 Motivation and Description

We extended the analysis stage, upon whose quality the whole pipeline relies, with a module for recovering non-local dependency relations (e.g., for control). Our approach is motivated by the recovery strategies in Levy and Manning (2004) . In contrast to that study, we perform our recovery over dependency graphs, instead of CFG trees.

The algorithm is a three-stage method: we first identify *loci* in the dependency graph, nodes which have a long-distance dependency to an antecedent node that is not directly represented in the graph. The second phase consists of determining what kind of relationship is missing (e.g., nominal subject). Finally, given the loci and relationship, we identify the most likely antecedent of the missing relation in the graph, and add the edge of the given type from the loci to the antecedent.

This is illustrated in Figure 2 for *Maler realized the importance of publishing his investigations* [derived from RTE4 955]. The word *publishing* is identified as a loci, and the missing relation identified is the nominal subject, the *nsubj* edge. We identify *Maler* as the likeliest antecedent, and insert the missing edge (dashed) from *publishing* to *Maler*.

Our model is learned from the annotations from the WSJ portion of the Penn Treebank to identify loci, their antecedents, and the edge type to recover in the corresponding dependency graphs for the parses. Classifiers are trained separately for each stage of the recovery and use graph neighborhood features to make confidence-weighted predictions for each of the

Dataset	Regular Acc	With Recovered
RTE2 dev	65.88	65.12
RTE2 test	61.25	63.38
RTE3 dev	70.12	71.75
RTE3 test	65.25	66.50
RTE4	62.60	62.70

Table 4: Effect of nonlocal recovery, performance on RTE 2, 3, and 4 (2-way Accuracy).

three variables. We choose the hypothesis with the highest combined score as antecedent. In addition, we use the Naive Hobbs pronominal recovery algorithm (Hobbs, 1986) to identify pronouns and antecedents. Using the same formulation, we remove the pronoun and replace the edge to its governor by an edge between its governor and its recovered antecedent.

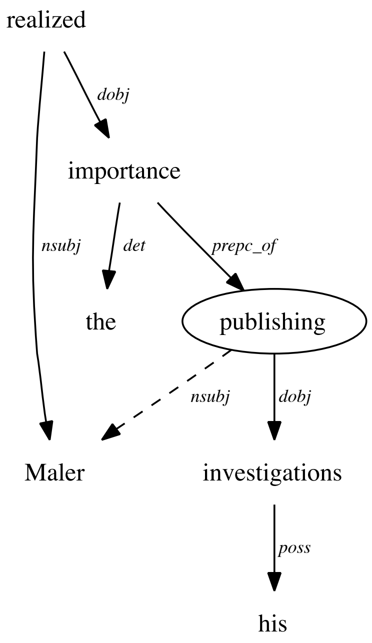

Figure 2: Dependency graph, loci circled, missing edge dashed.

4.2 Results and Analysis

In order to determine the contribution of edge recovery, we tested the Stochastic aligner using the regular dependency graphs versus those which had edges recovered via the above procedure over the datasets from RTE2, 3, and 4.

Results are given in Table 4. Note that this component was not active in the Stanford RTE-4 submissions. The numbers shown in Table 4 were produced with a more current version of the RTE system and are thus not directly comparable to Table 1.

Data analysis shows that the primary contribution from recovery is to make alignments easier, which in turn leads to a slight increase in the entailment score for certain problems. However, in most of the cases examined, the driving factor in an entailment decision was due to features in the inference stage, and not on edge recovery.

Acknowledgments. The authors acknowledge the support of the DARPA GALE program.

References

Brockett, Chris. 2007. Aligning the RTE 2006 Corpus. Tech. Rep. MSR-TR-2007-77, Microsoft Research.

Collins, Michael. 2002. Discriminative training methods for hidden Markov models. In *Proceedings of EMNLP*.

de Marneffe, Marie-Catherine, Trond Grenager, Bill MacCartney, Daniel Cer, Daniel Ramage, Chloé Kiddon, and Christopher D. Manning. 2007. Aligning semantic graphs for textual inference and machine reading. In *Proceedings of the AAAI Spring Symposium*.

de Marneffe, Marie-Catherine, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*.

de Marneffe, Marie-Catherine, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *Proceedings of ACL/HLT*.

Hobbs, J. 1986. Resolving pronoun references. In B. Grosz, K. Sparck-Jones, and B. Webber, eds., *Readings in Natural Language Processing*, pages 339–352. Morgan Kaufmann.

Klein, Dan and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL*.

Levy, Roger and Christopher D. Manning. 2004. Deep dependencies from context-free statistical parsers: Correcting the surface dependency approximation. In *Proceedings of ACL*, pages 327–334.

MacCartney, Bill, Michel Galley, and Christopher D. Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of EMNLP*.

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 NAACL*.

Ritter, Alan, Doug Downey, Stephen Soderland, and Oren Etzioni. 2008. It's a contradiciton – no, it's not: A case study using functional relations. In *Proceedings of EMNLP*.

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 on Recognising Textual Entailment*.