MITRE's Submissions to the EU Pascal RTE Challenge

**Samuel Bayer, John Burger, Lisa Ferro,
John Henderson, Alexander Yeh**

The MITRE Corporation
202 Burlington Rd.
Bedford, MA 01730, USA
{sam,john,lferro,jhndrsn,asy} @ mitre.org

Abstract

We describe MITRE's two submissions to the RTE Challenge, intended to exemplify two different ends of the spectrum of possibilities. The first submission is a traditional system based on linguistic analysis and inference, while the second is inspired by alignment approaches from machine translation. We also describe our efforts to build our own entailment corpus. Finally, we discuss our investigations and reflections on the strengths and weaknesses of the evaluation itself.

1 Background

The MITRE Corporation has a long-standing interest in both cutting-edge and practical approaches to text understanding. We believe that progress in task-independent text understanding requires an evaluation that pushes the research forward appropriately, and a substantial portion of our effort has been devoted to an in-depth exploration of using standard reading comprehension tests for this purpose (Hirschman et al., 1999). We have discovered, however, that the availability of such corpora is limited, their construction is expensive, and reading comprehension tests in general tend to be limited in their diagnostic ability.

In this context, the RTE (Recognizing Textual Entailment) Challenge appeals to us due to its generality, its simple structure, and the possibility that it might be significantly less expensive to develop the appropriate test corpora and sufficient training corpora, for those systems that require such. We also suspect that RTE techniques will be applicable to a broad range of problems.

For the challenge, MITRE developed two systems. We hypothesize that a successful RTE system will include elements of traditional approaches based on explicit linguistic analysis and inference, alongside robust, statistical approaches that leverage a range of simple, reliably extractable features. To clarify the shortcomings of each approach alone, and to help focus on how they might support each other, we implemented a system at each end of the continuum. System 1 is our traditional system, and System 2 is our statistical system.

2 System 1

System 1 is a baseline traditional system constructed using explicit modeling of linguistic analysis. The system processes both the Hypothesis and the Text using a MITRE-built tokenizer and sentence segmenter, the Ratnaparkhi (1996) POS tagger, the University of Sussex's Morph morphological analyzer (Minnon et al., 2001), the CMU Link Grammar parser (Sleator & Temperley, 1993), and a MITRE-built dependency analyzer and Davidsonian logic generator. The Text and Hypothesis are then compared using the University of Rochester's EPILOG event-oriented probabilistic inference engine (Schubert & Hwang, 2000). Very little additional semantic knowledge is exploited, beyond a few added inference rules and simple word lists for semantic classification. Due to its currently impoverished knowledge base, the system fails to prove entailment for virtually all of

the RTE data, and thus labels almost all of the data as non-entailing.

The results of System 1 on the test set are shown in Figure 1. Due to parse failures and other problems, the system failed to convert 213 of the 800 test pairs into the event logic, and so we made a partial submission for the other 587 test pairs. During development, pairs marked true were slightly more accurate than pairs marked false. This led us to a simplistic confidence scheme of 1.0 for true results and 0.5 for false results

System 1 currently has just two rules. One is intended to handle certain modals, e.g., *can run* does not entail *run*. This rule has no effect on the test set. The other rule handles some appositive cases. This other rule accounts for 2 of the correctly labeled Trues and 1 of the incorrectly labeled Trues.

Partly because System 1 has very few inference rules, about half of the correctly marked true pairs were pairs where the hypothesis is a simple subset of the text (e.g., *Rover is a big dog* entails *Rover is a dog*). However, this subset property of the inference engine also caused 6 of the 10 pairs incorrectly marked true; *Rover is not a dog* should not entail *Rover is a dog*, but System 1 thinks it does, due in part to our flat semantic representation (our modal rule was an attempt to address a small subset of these cases).

As we continue to work on this problem, we plan to exploit multiple potential sources of additional information: both explicit information sources like WordNet (Fellbaum, 1998) and information extracted from large background corpora such as Gigaword (Graff, 2003). We're also planning to synthesize this approach with the radically different approach found in System 2.

3 System 2

Statistical machine translation models inspire MITRE's second RTE system. These models are designed to find correspondences between pairs of sentences, and we believe that they can provide a stable starting point for capturing information needed to predict entailment. System 2 treats entailment data as an aligned translation corpus, and performs its prediction based on a combination of metrics intended to measure translation quality.

		System 1	System 2
Pairs processed		587	800
Correctly	T	11/285	231/400
Labeled	F	292/302	238/400
Accuracy		0.52	0.59
Precision		0.52	0.59
Recall		0.04	0.58
F-measure		0.07	0.58
CWS		0.50	0.62

Figure 1: System results

All but one of these metrics come from libparis, a library of string similarity metrics assembled by MITRE. Some of these metrics are inspired by MT evaluation, and some are standard string-matching algorithms (Gusfield, 1997). Additionally, we used an MT alignment score, on which we now focus our discussion.

Statistical MT explicitly models the probability that a sentence *F* in a source language will translate to a target language sentence *E*. Following Brown et al. (1993), most statistical MT models decompose this probability into many probabilities relating individual word-pairs in the two sentences. There are also mechanisms in the models for explaining spurious words in the source and target, which align with nothing.

Figure 2 shows an alignment example from the training data described below. We see that most of the source words either align with their identical counterparts or disappear. Additionally, *surrounded* aligns with *engulf*, and *Bushehr* with *Iran*. In general, we only hope that the MT models capture this sort of synonymy and paraphrase; we do not expect that these simple word associations can represent any complicated inference.

MT models must be trained from a corpus of *F-E* pairs, typically larger by orders of magnitude than the development set provided for the RTE evaluation. For this volume of data, we turned to the Gigaword newswire corpus (Graff, 2003), hypothesizing that newspaper headlines are often en-

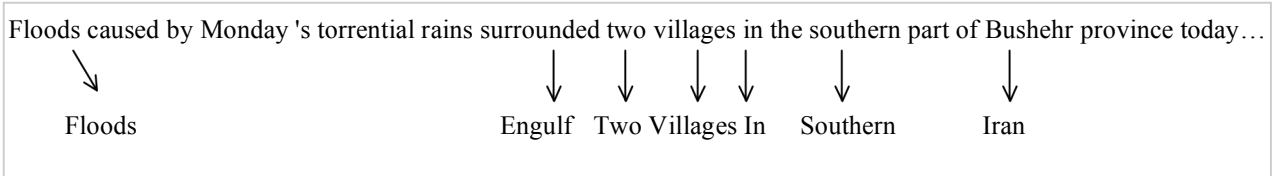

Figure 2: GIZA++ alignment for a training pair

differing	equal
heroism	gallantry
spaceflight	spacecraft
railmen	railworkers
procrastination	timing
hirsute	hair
engulf	surround
outplay	defeats
mountaineer	climber

Figure 3: A subset of the top word alignments acquired by GIZA++

tailed by the corresponding lead paragraph. We hoped that this noisy corpus would be a suitable training set for learning RTE alignments.

To test our hypothesis, we manually judged approximately 1000 of the lead-headline pairs from Gigaword for entailment (see Section 4 for a discussion of inter-annotator reliability). From this sample, we estimate that 60% of the headlines in the Gigaword corpus are entailed by the lead paragraph. We attempted to refine the data to acquire a smaller but less noisy corpus by training a document classifier (SVMlight: Joachims, 2002) to identify articles that exhibited the lead-entails-headline quality. Like those classifiers used to predict genre or topic, this training included the entire articles with bag-of-words features. We experimented with active learning techniques, and finally derived a 100,000-document subset of Gigaword with approximately 75% lead-entails-headline purity.

We used the GIZA++ toolkit (Och & Ney, 2003) to induce alignment models on the paired leads and headlines from the Gigaword subset. Some indicative word correspondences found by the model are shown in Figure 3. When applied to our (held-out) manually judged Gigaword data, these models could predict headline entailment with roughly 80% accuracy (compared to the base rate of 60% in that development set).

Unfortunately the alignment scores alone were next to useless for the RTE development data, predicting entailment correctly only slightly above chance. This is presumably because the negative instances in the RTE data are designed to have substantial conceptual overlap between the text and hypothesis, while the negative Gigaword instances frequently have little overlap.

At this point, we combined the alignment models with the libparis metrics described earlier. We first trained an SVM classifier on the RTE development data, using these features, but cross-validation experiments showed this to be unpromising as well—the data appeared to be far from linearly separable. In the end, we combined all the features using a simple *k*-nearest-neighbor classifier that chose, for each test pair, the dominant truth value among the five nearest neighbors in the development set. Results are shown in Figure 1.

4 The Corpus and the Evaluation

The RTE evaluation, while promising, faces a number of challenges as it matures.

First is the issue of the feasibility of the task. Based on our investigations, the task appears to be quite difficult for humans. When tested on 10 pairs from each of the seven application scenarios in the dev2 training set, our human judge achieved an agreement rate of 91% (64/70) compared to the given truth values. While this number might seem impressive, it is less so when one considers that the training data was already considerably simplified from a real-world application. According to the Task Definition, the *T-H* pairs were hand-crafted, and any pairs "for which there was disagreement among the judges were discarded." Thus, the 91% agreement is somewhat troubling.

We also attempted to determine the degree to which paraphrases played a role. Two of our researchers independently reviewed all the TRUE entailments of the dev2 set, and determined that 94% (131/140) were mere paraphrases (*John murdered Bill* → *Bill was killed by John*), as opposed to classic entailments (*Bill is dead*). During this process, we uncovered many cases where we disagreed with the given truth value on the grounds of synonymy (e.g., *in bloody clothes* → *covered in blood*). We also identified potential disagreements about the extent to which world knowledge is allowed to play a role. For instance, pair 102 (*domestic threat* → *threat of attack*) is more convincing if one understands the implications of *al Qaeda* and *September 11, 2001* mentioned in the text.

In the process of building our own training corpus (see Section 3), we conducted additional inter-judge studies. Even after one trial phase and with a supplementary set of guidelines in hand, the judges achieved only 81% inter-annotator agreement. While a portion of this disagreement is due

to the messiness of the data (e.g., bylines and date lines mis-zoned into the headlines), the more egregious difficulty was that our judges found they had irreconcilable differences in meaning interpretation. For example, in the following lead-headline pair, one judge did not think that *safe operation* entailed (meant the same thing as) *operates smoothly*, and one did.

- *As of Saturday, Shanghai's Hongqiao Airport has performed safe operation for some 2,600 consecutive days, setting a record in the country*.
- *Shanghai's Hongqiao Airport Operates Smoothly*

It's hard to imagine how annotation guidelines would resolve this disagreement. This leads us, obviously, to wonder how an evaluation like this might be designed to ensure more consistent human judgment. It also suggests that if the organizers pre-clean the development corpus in future RTE evaluations as they did for this evaluation, it would be quite useful for them to report the percentage of pairs eliminated.

In addition to the challenges of interannotator agreement, it isn't clear what a "representative" corpus would look like. The RTE development corpus is clearly constructed to stress-test a range of legitimate and illegitimate inferences, but it is not clear how to balance these. It is unclear exactly how this technology will be used, and so it is equally unclear which issues might be more vs. less important to represent in an evaluation. Even in the cases where RTE data has been drawn from "naturally occurring" corpora, such as multiple, parallel translations, it's unclear how RTE technology would be applied to those corpora.

5 Conclusion and Future Work

It's been said, about difficult challenges like RTE, that one should be aware of the temptation to climb a tree in order to get to the moon (Dreyfus, 1979); i.e., short-term solutions can be initially superior, but are frequently dead ends. MITRE's two entries illustrate this dilemma quite clearly. System 1 is the rocket ship with nothing inside: fiendishly difficult to get off the ground, and unable to fly until a wide number of things work fairly well. System 2, on the other hand, is a tree. Our challenge, as we move forward, is to figure out how to leverage the strengths and potential of both.

Acknowledgments

This paper reports on work supported by the MITRE Sponsored Research Program. We would also like to extend our thanks to Fabrizio Morbini at the University of Rochester, without whose diligent support of EPILOG this work would not have been possible.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer, 1993. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2).

Hubert Dreyfus, 1979. *What Computers Can't Do: A Critique Of Artificial Reason*. Harper and Row.

Christiane Fellbaum, 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

David Graff, 2003. *English Gigaword*. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05

Dan Gusfield, 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press.

Lynette Hirschman, Marc Light, Eric Breck and John D. Burger, 1999. Deep Read: A reading comprehension system. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*.

Thorsten Joachims, 2002. *Learning to Classify Text Using Support Vector Machines*. Kluwer.

Guido Minnen, John Carroll and Darren Pearce, 2001. Applied morphological processing of English. *Natural Language Engineering,* 7(3).

Franz Josef Och and Hermann Ney, 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1)

Adwait Ratnaparkhi, 1996. A maximum entropy part-of-speech tagger. *Proceedings of the Empirical Methods in Natural Language Processing Conference.*

Lenhart K. Schubert and Chung Hee Hwang, 2000. Episodic Logic meets Little Red Riding Hood: A comprehensive, natural representation for language understanding, in L. Iwanska and S.C. Shapiro (eds.), *Natural Language Processing and Knowledge Representation*. MIT/AAAI Press.

Daniel Sleator and Davy Temperley, 1993. Parsing English with a link grammar. *Third International Workshop on Parsing Technologies*.

Ben Wellner, Lisa Ferro, Warren Greiff and Lynette Hirschman, 2005. Evaluating language understanding through reading comprehension tests. *Natural Language Engineering* (to appear).