applying them only if their L part was matched in $\mathcal{F}$ and their R part was matched in $\mathcal{H}$.

The system was trained over the RTE-3 development set, and tested on both RTE-3 test set and RTE-4 (which includes only a test set).

Results Table 2 provides statistics on rule application using all rule bases, over the RTE-3 development set and the RTE-4 dataset[6]. Overall, the primary result is that the compact forest indeed accommodates well extensive rule application from large-scale rule bases. The resulting forest size is kept small, even in the maximal cases which were causing memory overflow for explicit inference.

The accuracies obtained in this experiment and the overall contribution of rule-based inference are shown in Table 3. The results on RTE-3 are quite competitive: compared to our 66.4%, only 3 teams out of the 26 who participated in RTE-3 scored higher than 67%, and three more systems scored between 66% and 67%. The results for RTE4 rank 9-10 out of 26, with only 6 teams scoring higher by more than 1%. Overall, these results validate that the setting of our experiment represents a state-of-the-art system.

Inference over the rule bases utilized in our experiment improved the accuracy on both test sets. The contribution was more prominent for the RTE-4 dataset. These results illustrate a typical contribution of current knowledge sources for current RTE systems. This contribution is likely to increase with current and near future research, on topics such as extending and improving knowledge resources, applying them only in semantically suitable contexts, improved classification features and broader search strategies. As for our current experiment, we may conclude that the goal of assessing the compact forest scalability in a state-of-the-art setting was achieved [7].

Finally, Tables 4 and 5 illustrate the usage and contribution of individual rule bases. Table 4 shows the distribution of rule applications over the various rule bases. Table 5 presents ablation study showing the marginal accuracy gain for each rule base. These results show that each of the rule bases is applicable for a large portion of the pairs, and contributes to the overall accuracy.

[6]Running time is omitted since most of it was dedicated to rule fetching, which was rather slow for our available implementation of some resources. The elapsed time was a few seconds per (t, h) pair.

[7]We note that common RTE research issues, such as improving accuracy, fall out of the scope of the current paper.

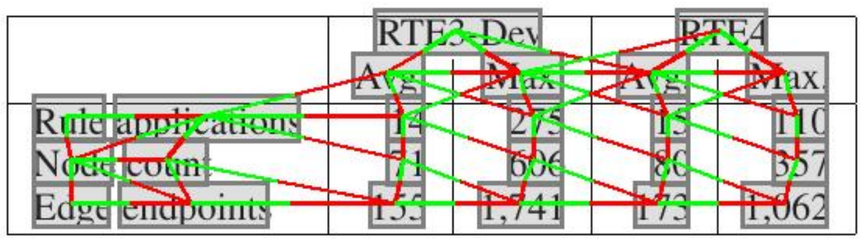

Table 2: Application of compact inference to the RTE-3 Dev. and RTE-4 datasets, using all rule types.

Test set	Accuracy: No inference	Accuracy: Inference	Δ
RTE3	64.6%	66.4%	1.8%
RTE4	57.5%	60.6%	*3.1%

Table 3: Inference contribution to RTE performance. The system was trained on the RTE-3 development set. * indicates statistically significant difference (at level $p < 0.02$, using McNemar's test).

Rule base	RTE3-Dev Rules	RTE3-Dev App	RTE4 Rules	RTE4 App
WordNet	0.6	1.2	0.6	1.1
AmWN	0.3	0.4	0.3	0.4
Wikipedia	0.6	1.7	0.6	1.3
DIRT	0.5	0.7	0.5	1.0
Generic	4.7	10.4	5.4	11.5
Polarity	0.2	0.2	0.2	0.2

Table 4: Average number of rule applications per (t, h) pair, for each rule base. *App* counts each rule application, while *Rules* ignores multiple matches of the same rule in the same iteration.

Rule base	ΔAccuracy (RTE4)
WordNet	0.8%
AmWN	0.7%
Wikipedia	1.0%
DIRT	0.9%
Generic	0.4%
Polarity	0.9%

Table 5: Contribution of various rule bases. Results show accuracy loss on RTE-4, obtained for removing each rule base (ablation tests).

5 Related Work

This section discusses related work on applying knowledge-based transformations within RTE systems, as well as on using packed representations in other NLP tasks.

RTE Systems Previous RTE systems usually restricted both the type of allowed transformations and the search space. Systems based on lexical (word-based or phrase-based) matching of h in t typically applied only lexical rules (without vari-