Analysis Report - Assignment 3

Ilay Amsalem
Rom Widergoren

1 Experimental Setup

We evaluated our implementation of the DIRT similarity measure on the provided test set of predicate pairs, consisting of positive (entailing) and negative (non-entailing) examples.

Experiments were conducted using two corpus sizes:

- **Small input**: 10 syntactic-ngram files
- **Large input**: 100 syntactic-ngram files

The data was taken from the Google Syntactic N-grams corpus (biarcs collection). For each predicate pair, a similarity score was computed based on the overlap of slot fillers, weighted by mutual information, following the DIRT framework.

2 Threshold Selection

Similarity scores were converted into binary entailment decisions using a threshold-based rule:

$$\text{similarity} \geq \tau \Rightarrow \text{predicted entailment}$$

$$\text{similarity} < \tau \Rightarrow \text{predicted non-entailment}$$

Thresholds were selected empirically by evaluating Precision, Recall, and F1-measure at multiple similarity cutoffs.

2.1 Threshold for Small Input

Similarity scores obtained from the small input were substantially lower due to severe data sparsity. The small test set contains only positive (entailing) predicate pairs, with no negative examples.

Although very low thresholds yield perfect F1 by trivially predicting all pairs as entailments, such thresholds provide no meaningful filtering. We therefore selected:

$$\tau = 0.06$$

This threshold removes extremely weak similarities while preserving a majority of true entailments.

2.2 Threshold for Large Input

For the large input, system performance was evaluated at multiple similarity thresholds in the range $[0.0, 0.4]$. The test set is highly imbalanced, containing 481 positive examples and only 19 negative examples.

The threshold that maximizes the F1-measure is:

$$\tau = 0.04$$

This value provides the best tradeoff between precision and recall.

3 Evaluation Metrics

System performance was evaluated using Precision, Recall, and F1-measure, computed from the binary entailment predictions induced by the selected thresholds.

3.1 Large Input (100 files)

Using a similarity threshold of $\tau = 0.04$, the system produced the following confusion matrix counts:

- True Positives (TP): 429
- False Positives (FP): 13
- False Negatives (FN): 52
- True Negatives (TN): 6

From these counts, we obtain:

$$\text{Precision} = \frac{TP}{TP + FP} = 0.971$$

$$\text{Recall} = \frac{TP}{TP + FN} = 0.892$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = 0.930$$

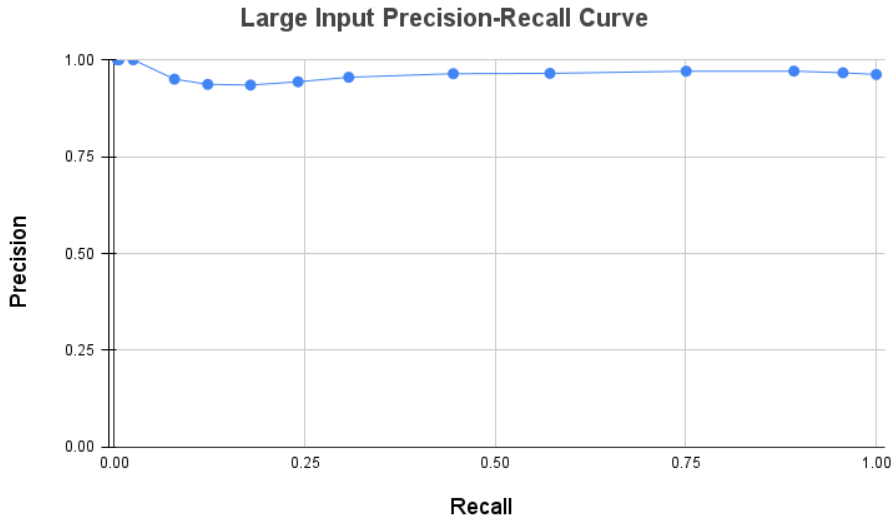

Figure 1: Precision–Recall curve for the large input setting, obtained by evaluating the system at multiple similarity thresholds.

Precision remains high across most thresholds due to the small number of negative examples, while recall decreases steadily as the threshold increases. The selected operating point at $\tau = 0.04$ achieves the highest F1-score.

3.2 Small Input (10 files)

Using a similarity threshold of $\tau = 0.06$, the system was evaluated on the small input test set, which consists exclusively of positive (entailing) predicate pairs.

The resulting confusion matrix is:

- True Positives (TP): 9
- False Positives (FP): 0
- False Negatives (FN): 5
- True Negatives (TN): 0

This yields the following evaluation scores:

$$\text{Precision} = 1.000 \quad \text{Recall} = 0.643 \quad \text{F1} = 0.783$$

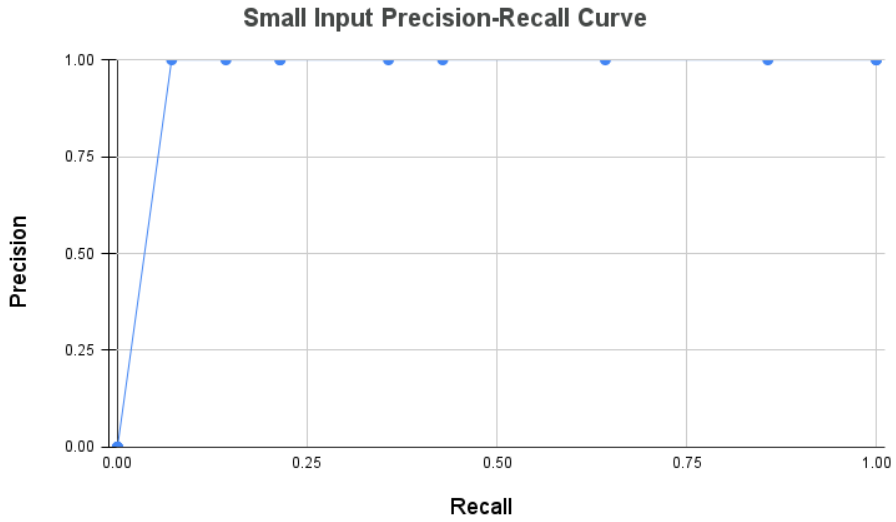

Figure 2: Precision–Recall curve for the small input setting.

Since the small test set contains no negative examples, precision remains perfect for all thresholds at which at least one predicate pair is predicted as an entailment. Recall decreases monotonically as the similarity threshold increases, reflecting data sparsity and limited slot overlap.

4 Error Analysis (Large Input)

Error analysis was performed using the selected threshold $\tau = 0.04$.

4.1 True Positives (TP)

Predicate 1	Predicate 2	Score
X lead to Y	X result in Y	0.3801
X associate with Y	X occur with Y	0.2398
X protect against Y	X protect from Y	0.3853
X metabol by Y	X metabol in Y	0.2762
X transmit by Y	X transmit through Y	0.2785

These examples correspond to near-paraphrases or strong entailment relations with highly overlapping slot fillers.

4.2 False Positives (FP)

Predicate 1	Predicate 2	Score
X constrict Y	X dilate Y	0.2131
X confound with Y	X distinguish from Y	0.1644
X associate with Y	X distinguish from Y	0.1860
X prevent Y	X produce Y	0.1745
X differ from Y	X resemble Y	0.2088

False positives typically arise from high topical relatedness despite contradictory or non-entailing semantics, reflecting the lack of polarity modeling.

4.3 False Negatives (FN)

Predicate 1	Predicate 2	Score
X discover Y	X invent Y	0.00053
X develop Y	X produce Y	0.00080
X introduce Y	X make Y	0.00062
X attack Y	X destroy Y	0.00634
X die of Y	X get Y	0.02606

These entailments receive low similarity scores due to sparse or abstract slot distributions.

4.4 True Negatives (TN)

Predicate 1	Predicate 2	Score
X derive from Y	X destroy Y	0.00177
X be in Y	X have Y	0.04313
X know as Y	X resemble Y	0.04099
X convert to Y	X derive from Y	0.08454
X die of Y	X get Y	0.02606

4.5 Small vs. Large Input Comparison

Comparing the same predicate pairs across the small and large input settings reveals a consistent trend. For most true-positive examples, similarity scores increase substantially in the large input due to improved coverage of dependency paths and more reliable mutual information estimates. Conversely, many false negatives in the small input receive extremely low similarity scores that are partially corrected in the large input.

This comparison highlights the impact of data sparsity on DIRT-style similarity measures and explains the significant improvement in recall observed when moving from the small to the large corpus.

5 Summary

The analysis demonstrates that the system successfully captures many meaningful entailment relations, particularly paraphrases and closely related predicates. Increasing the corpus size substantially improves recall while maintaining high precision, even in the presence of a highly imbalanced test set.

Remaining errors primarily stem from antonymy, polarity reversal, and rare or abstract predicates. Overall, the observed behavior aligns well with known properties and limitations of the DIRT algorithm.