

Assignment 3 — DIRT: Extraction of Lexico-syntactic Similarities

Analysis Report

1. Experimental Setup

Small run

- Input size: **10 Biarcs files**
- Outputs produced:
 - **mi.tsv**
 - **out_scores.tsv** (scores for all test pairs)

Large run (placeholder)

- Input size: **100 Biarcs files** (or as many as budget allows)
- Outputs to be produced:
 - **mi.tsv** (large)
 - **out_scores.tsv** (large)

Expected differences (Small vs Large):

- **Coverage:** large input should yield many more predicates and fillers, reducing sparsity.
 - **Score distribution:** more non-zero similarities, better threshold separation, but also potentially more noise and false positives.
 - **Evaluation stability:** metrics on 10 files can be unstable due to many zero scores.
-

2. Small Run Results (10 files)

2.1 Dataset sizes

- Positive pairs: 2481
- Negative pairs: 99
- Total pairs scored: 2580

2.2 Score distribution observation (Small)

On the 10-file run, the similarity scores are extremely sparse:

- Most pairs receive **score = 0.0**.
- In this run, **all negative pairs received score = 0.0**, and only a small number of positive pairs received a score > 0.

This behavior is consistent with feature sparsity:

- predicates may not be found in the small corpus, or
- the predicates exist but have no shared high-MI fillers in either slot, producing zero similarity.

2.3 Choosing a threshold and F1 (Small)

We define the classifier:

- predict **positive** if `score >= threshold`
- predict **negative** otherwise

On this run:

- The threshold that maximizes F1 is **threshold = 0.0**, which predicts all pairs as positive.
 - This yields very high F1 mainly because the test set is heavily imbalanced toward positives.

Small run metrics (threshold = 0.0)

- TP = 2481, FP = 99, TN = 0, FN = 0
- Precision = 0.9616
- Recall = 1.0000
- F1 = 0.9804

Interpretation This threshold is not informative (it does not separate classes). Therefore, for diagnostic purposes we also examine a “strict” threshold:

- **threshold > 0** (e.g., `1e-9`) treats `score=0` as negative.

Small run diagnostic metrics (threshold = 1e-9)

- TP = 9, FP = 0, TN = 99, FN = 2472
- Precision = 1.0000
- Recall = 0.00363
- F1 = 0.00723

This indicates that on 10 files the model produces non-zero similarity for very few positive pairs, and none for negatives.

3. Precision–Recall Curve (Small)

A precision–recall curve is produced by sweeping the threshold over the set of observed scores.

Small run PR behavior

- At threshold 0.0: recall is 1.0 and precision ≈ 0.962 (predict all positives).
- For any threshold > 0: precision becomes 1.0 but recall collapses (only a handful of pairs remain predicted positive).

(Insert figure here: PR curve for Small)

4. Error Analysis (Small)

4.1 Threshold used for error categorization

For error categorization (TP/FP/TN/FN) we use **threshold = 1e-9** (i.e., “score must be > 0”).

Reason: threshold 0.0 yields no TN/FN, making error analysis impossible.

4.2 Category counts (Small, threshold = 1e-9)

- TP: 9
- FP: 0
- TN: 99
- FN: 2472

Note: **No false positives exist in the small run** because all negative pairs scored 0. This may change on the large run once negatives start receiving non-zero similarity due to broader feature overlap.

4.3 Examples (Small)

True Positives (top examples)

1. X attack Y ↔ X affect Y (0.0299)
2. X accommodate Y ↔ X accommodate by Y (0.1869)
3. X associate with Y ↔ X accompany by Y (0.1769)
4. X accompany with Y ↔ X accompany by Y (0.1413)
5. X attend with Y ↔ X accompany by Y (0.1385)

True Negatives (examples; all had score 0.0)

1. X ... ↔ Y ... (0.0)
2. X ... ↔ Y ... (0.0)
3. X ... ↔ Y ... (0.0)
4. X ... ↔ Y ... (0.0)
5. X ... ↔ Y ... (0.0)

(Replace “...” with concrete pairs from your `out_scores.tsv`—see code section below that prints examples automatically.)

False Negatives (examples; positives with score 0.0)

1. (pos) p1 ↔ p2 (0.0)
2. (pos) p1 ↔ p2 (0.0)
3. (pos) p1 ↔ p2 (0.0)
4. (pos) p1 ↔ p2 (0.0)
5. (pos) p1 ↔ p2 (0.0)

False Positives

- None observed in Small run at threshold > 0.

4.4 Common behaviors (Small)

- **Sparsity dominates:** most similarities are zero because shared MI-weighted fillers are rare in 10 files.
- **Coverage gaps:** some test predicates may not appear in the small corpus at all, producing empty feature vectors.

- **Large run expectation:** as the corpus grows, more predicates and fillers appear, increasing overlap; this should raise recall at meaningful thresholds, but may introduce FPs.
-

5. Large Run Placeholders (100 files)

Repeat Sections 2–4 with the large outputs:

- Recompute threshold choice from the large score distribution.
- Produce PR curve for large.
- Provide 5 examples each of TP/FP/TN/FN and compare their scores to the small run.

Expected difference

- Large run should produce more non-zero scores for both positives and negatives, making PR curves and thresholding substantially more informative than the small run.