Transparent Rationale Generator (TRG)
Pseudocode, Formulation, and Explanations

1 Notation and Glossary

Symbol	Meaning
$S = [s_1, \ldots, s_L]$	Input sentence tokens
$h_j \in \mathbb{R}^d$	Contextual embedding of token j
w_j	Type (word/subword group) of token j
$C_w = \{c_{w,1}, \ldots, c_{w,K_w}\}$	Prototype centroids for type w
$\mathrm{sim}_{j,i}$	Cosine similarity between h_j and prototype $c_{w,i}$
p_j	Softmax distribution over senses for token j
p_{max}	Maximum probability in p_j
$\hat{y}$	Argmax sense index (chosen sense)
$H(p_j)$	Entropy of distribution p_j
Var_{p_j}	MC-dropout variance across predictions
σ_j	Aleatoric uncertainty predicted by σ-Net
d_j^{min}	Distance to nearest prototype
U_j	Normalized aggregated uncertainty score
g_j	Gate indicating whether token needs explanation
E_j	Extracted evidence tokens
ASBN_diag_j	Diagnostic signals from ASBN discriminators
G	Generator model (structured input → rationale text)
V	Verifier model (source + rationale → sense distribution)

2 Mathematical Formulation

2.1 Prototype-based sense distribution

$$\mathrm{sim}_{j,i} = \cos(h_j, c_{w,i}), \quad p_j = \mathrm{softmax}\left(\frac{\mathrm{sim}_j}{T}\right)$$

2.2 Entropy (normalized)

$$H_j = -\sum_i p_j[i] \log(p_j[i] + \epsilon), \quad H_{\text{norm}} = \frac{H_j}{\log K_w}$$

2.3 MC-dropout variance (epistemic)

$$\mu_p = \frac{1}{M} \sum_m p_j^{(m)}, \quad \mathrm{Var}_{p_j} = \frac{1}{M} \sum_m \|p_j^{(m)} - \mu_p\|^2$$

2.4 Aleatoric uncertainty

$$\text{aleatoric}_j = e^{s_j}, \quad \text{aleatoric}_{\text{norm}} = \frac{\text{aleatoric}_j}{\text{aleatoric}_j + c}$$

2.5 Novelty to prototypes

$$d_j^{\text{min}} = \min_i (1 - \text{sim}_{j,i}), \quad d_j^{\text{norm}} = \sigma\left(\gamma \cdot \frac{d_j^{\text{min}} - \mu_d}{\sigma_d + \epsilon}\right)$$

2.6 Aggregate uncertainty

$$U_j = \sigma\Big(\kappa \cdot \Big(\alpha_H H_{\text{norm}} + \alpha_{\text{var}} \text{Var}_{p_j} + \alpha_a \text{aleatoric}_{\text{norm}} + \alpha_d d_j^{\text{norm}} - \tau_U\Big)\Big)$$

3 TRG Training Algorithm (Silver-Only)

Algorithm 1 TRG Training with Silver Rationales

```
Input: Training corpus, DSCD+ASBN outputs
Output: Generator $G$, Verifier $V$
Run DSCD+ASBN to compute $h_j, p_j, U_j, E_j$, ASBN_diag$_j$
for each token $j$ flagged (if $U_j > \tau_U$) do
    Extract evidence tokens $E_j$ from attention
    Select top alternative senses Alts$_j$
    Collect prototype examples ProtoEx$_j$
    Summarize audit diagnostics Audit$_j$
    Format structured input $X_j$
    Build silver rationale $R_j = \text{Template}(X_j)$
    Add $(X_j, R_j, \hat{y})$ to training set
end for
Initialize $G, V$
for epoch = 1 to $E$ do
    for batch $(X, R, y)$ do
        $L_{\text{gen}} \leftarrow$ CrossEntropy($G(X), R$)
        $q \leftarrow V(S \oplus R)$
        $L_{\text{fid}} \leftarrow$ CrossEntropy($q, y$)
        $L_{\text{cov}} \leftarrow$ CoverageLoss($G(X), E$)
        $L \leftarrow L_{\text{gen}} + \lambda_{\text{fid}} L_{\text{fid}} + \lambda_{\text{cov}} L_{\text{cov}}$
        Update $G, V$ by backpropagation
    end for
end for
```

4 TRG Inference Algorithm

Algorithm 2 TRG Inference (Runtime)

```
Input: Sentence $S$, DSCD outputs
for each token $j$ do
  if $p_{\max} > \tau_{\text{high}}$ and $U_j < \tau_{\text{low}}$ then
    $R_j \leftarrow \text{Template}(X_j)$
  else
    Generate candidates $R^c = G(X_j)$
    for each $r \in R^c$ do
      $q = V(S \oplus r)$
      if $\arg\max q = \hat{y}$ and $q[\hat{y}] >$ threshold then
        $R_j \leftarrow r$, break
      end if
    end for
    if no candidate accepted then
      $R_j \leftarrow \text{Template}(X_j)$
    end if
  end if
  Output $\{j, \hat{y}, p_{\max}, E_j, R_j, \text{Alts}_j, \text{ASBN\_diag}_j, U_j\}$
end for
```

5 Detailed Explanation of Components

5.1 Extractor

The extractor collects signals from DSCD and ASBN for each token. It includes contextual embeddings, probability distributions over senses, uncertainty estimates, evidence tokens from attention, prototype assignments, and diagnostic notes from discriminators. This ensures rationales are grounded in model internals.

5.2 Formatter

The formatter converts raw extractor outputs into structured fields such as the token, chosen sense, confidence score, alternatives, evidence tokens, prototype examples, and ASBN diagnostics. This structured format enables reliable template generation and safe input to the generator.

5.3 Template

The template is a deterministic slot-filler that produces a concise rationale from structured inputs. For example:

> Chose “page” (conf 0.86) because বইয়ের indicates book context. Alternatives “leaf” (0.12) and “blade” (0.02) are less likely. No ASBN bias detected.

Templates are always faithful to model signals and serve as fallback explanations when the generator is uncertain.

5.4 Generator (G)

The generator is a small seq2seq model trained on silver rationales. It rewrites structured inputs or template outputs into more natural explanations. This improves fluency while maintaining faithfulness to DSCD+ASBN signals.

5.5 Verifier (V)

The verifier is a classifier that checks whether a rationale is faithful. It predicts the sense given the sentence and rationale. If it agrees with the DSCD prediction, the rationale is accepted. This prevents fluent but misleading explanations.

5.6 Loss Terms

- L_{gen}: Cross-entropy between generator outputs and silver rationales (fluency).
- L_{fid}: Cross-entropy on verifier predictions (faithfulness).
- L_{cov}: Penalty if evidence tokens are absent in generated rationale (coverage).
- Total loss:

$$L = L_{\text{gen}} + \lambda_{\text{fid}} L_{\text{fid}} + \lambda_{\text{cov}} L_{\text{cov}}$$

5.7 Training Phase

During training, tokens with high uncertainty are selected. Silver rationales are generated using templates. Structured inputs and rationales form the training set. The generator and verifier are trained jointly with combined loss, ensuring both fluency and faithfulness.

5.8 Inference Phase

At runtime, if confidence is high and uncertainty is low, a template rationale is used. Otherwise, the generator produces candidate rationales, which are verified for faithfulness. If verification fails, fallback to template occurs. The output includes the rationale text, evidence, chosen sense, alternatives, ASBN diagnostics, and uncertainty.

5.9 Example Walkthrough

For the sentence: "সে বইয়ের পাতায় ছবি আঁকছে।", the token "পাতা" has senses [leaf:0.12, page:0.86, blade:0.02]. Evidence token = "বইয়ের". The template rationale is:

> Chose “page” (conf 0.86) because বইয়ের indicates book context. Alternatives ”leaf” (0.12) and “blade” (0.02) are less likely.

The generator may produce a more fluent explanation if accepted by the verifier.