RST Signals as Features Parser Error/Success Prediction Tool

Funded by the European Union

UK Research and Innovation

Usage/License

Download

GitHub repository: https://github.com/metabolean5/signals-as-features

Researchers

- Martial Pastor, Radboud University
- Nelleke Oostdijk, Radboud University

Centre for Language Studies, The Netherlands

Summary

A signal-aware diagnostic framework for RST discourse parsing. Given a coherence relation and its associated linguistic signals, the tool predicts whether a parser will succeed or fail on that relation — enabling targeted error analysis and interpretable parser diagnostics without retraining.

What You Can Do With It

- Parser Diagnostics: Run DMRST parser outputs through the pipeline to identify success/failure patterns by relation and signal type.
- Signal Auditing: Align RST Signaling Corpus annotations with RST-DT-compatible parser outputs and examine which signals are associated with parsing errors.
- Feature-Based Error Analysis: Use the binary signal vectors in your own classifiers or analysis pipelines to study how linguistic signals affect parsing difficulty.
- Distractor Detection: Identify discourse markers and lexical signals that consistently predict parsing errors, supporting targeted data augmentation and parser fine-tuning.

How It Works

- **Step 1 — Signal Alignment**: Align RST Signaling Corpus annotations with parser outputs at token level using the corrected alignment script included in the repository.
- **Step 2 — Feature Encoding:** Encode each relation as a binary vector over 50 signal types (discourse markers, lexical, syntactic, semantic, reference, etc.).
- **Step 3 — Prediction:** An XGBoost classifier predicts parser performance:

Input: Binary signal vector

Output: {1 = SUCCESS, 0 = ERROR}

The repository includes a trained model and supports retraining on any RST-DT-compatible parser.

What We Offer

- Alignment Code: Token-level alignment of RST Signaling Corpus annotations to RST-DT parser outputs, with corrected position calculation.
- Signal Encoder: Relation-to-binary-vector encoding across all 50 signal types in the Das & Taboada taxonomy.
- Trained Classifier: XGBoost model checkpoint for drop-in error/success prediction on new parser outputs.
- Annotated Data: 2,306 DMRST-parsed relations with gold signal annotations and error/success labels, ready for downstream analysis.

hybrids

HYBRIDS Website

www.hybridsproject.eu

Get in Touch

martial.pastor@ru.nl