TECHNISCHE UNIVERSITÄT WIEN
Vienna | Austria

Topic 8 - Explainable Relation Extraction

Natural Language Processing and Information Extraction 2025W

Group: Token 13

Ege Aydin - 12432147
Hasan Berke Bankoglu - 12432802
Ege Özbaran - 12433722
Muhammad Bilal Hussain - 12442081

January 16, 2026

Table of Contents

Introduction: Relation Extraction

MS1

What is Relation Extraction?

- Entities in text
- Semantic relations between entities
- Relation types or patterns
- Often directional

"The Burst was caused by pressure" ⟶ Cause-Effect(e2,e1)

Introduction: Relation Extraction

MS1

Why Relation Extraction Matters

- Structured knowledge
- Industry NLP task
- Black-box models

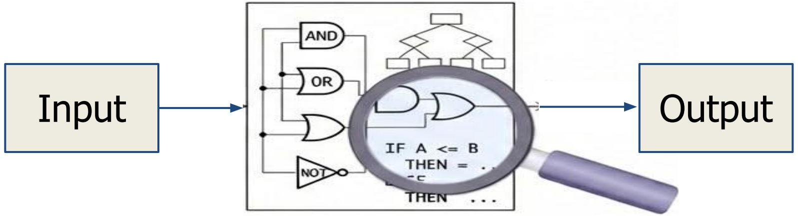
Introduction: Explainability Challenge

MS1

Machine Learning Based Systems

- High predictive performance
- Data-driven learning
- Generalize well across domains
- Limited interpretability
- Black-box decision process

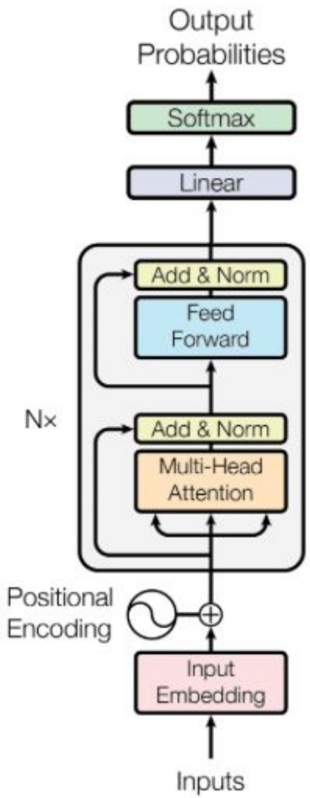
Rule Based Systems



- Fully interpretable
- High precision
- Deterministic behavior
- Limited scalability

Black-Box Model: BERT

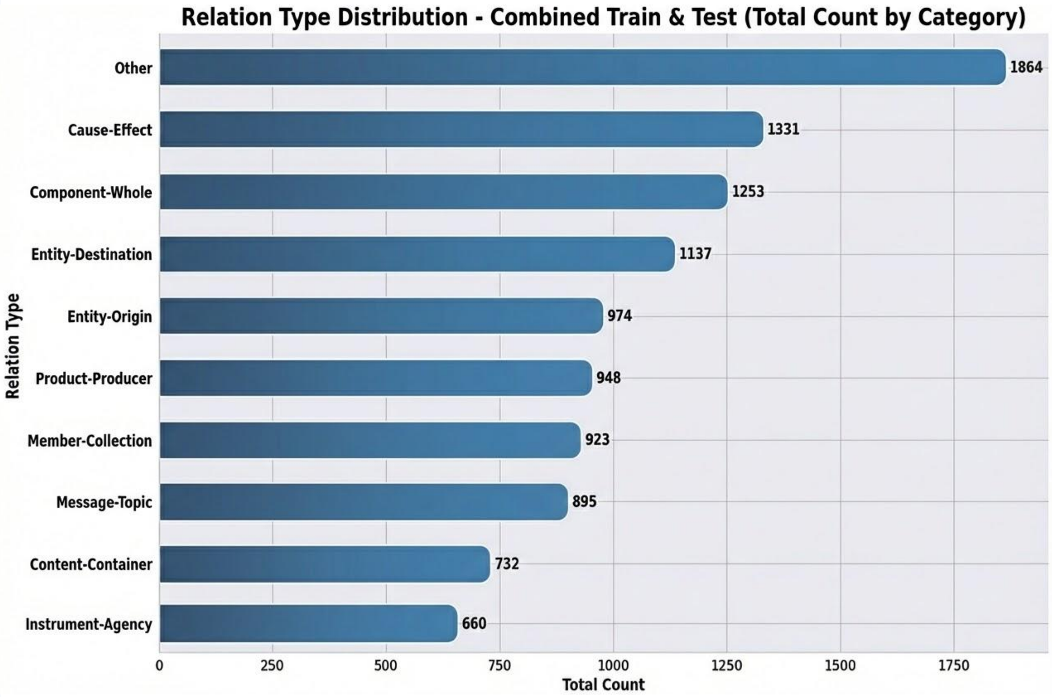
MS2



- Transformer encoder–based language model
- Learns bidirectional context
- Pretrained model using Masked Language Modeling (MLM)
- Produces contextual token embeddings
- RoBERTa

Introduction: Dataset & Task

MS1



SemEval-2010 Task 8

- Standard relation extraction benchmark
- Sentence-level relation classification
- Two marked entities (e_1, e_2) per sentence (NER done)
- 9 directed relation types + other
- In total: 10717 sentences (8000 + 2717)

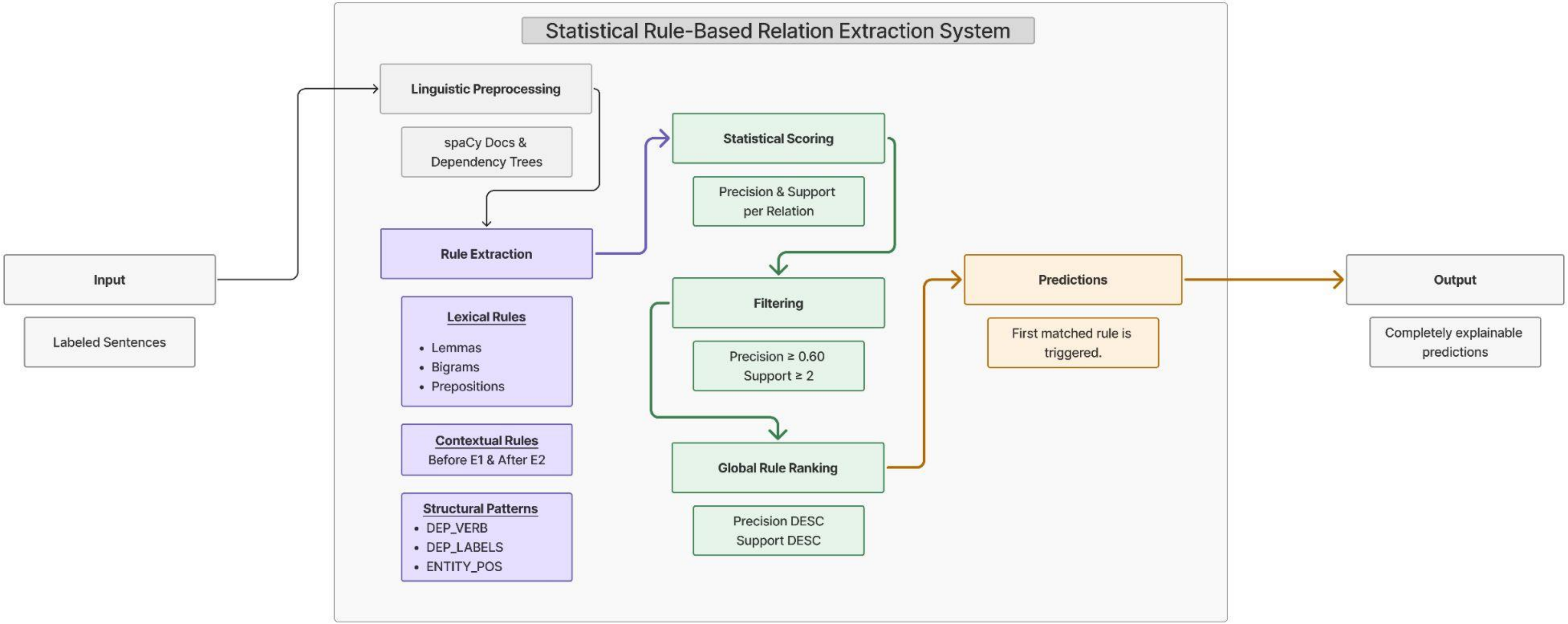
Approach: Rule Based System

MS2

Statistical Rule Based Relation Extraction System

Approach: Rule Based System

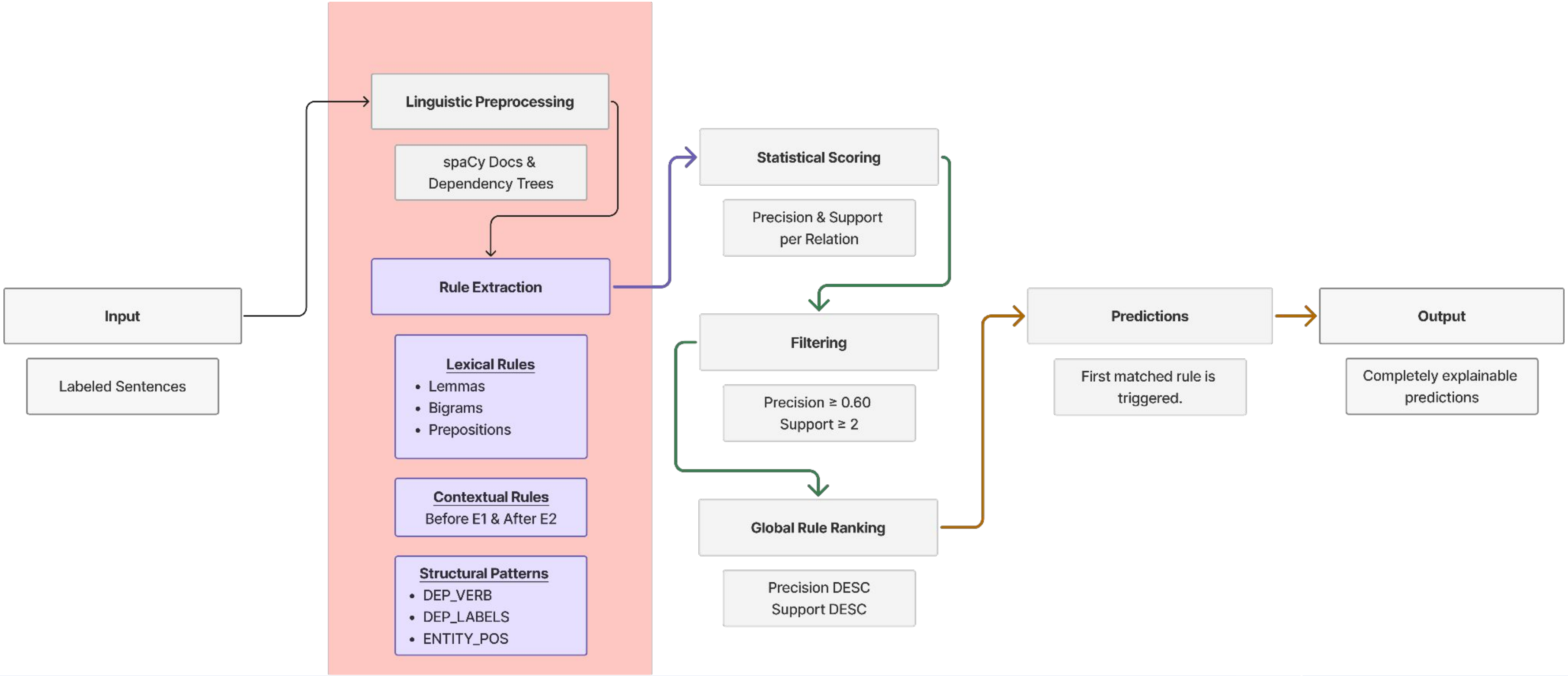
MS2



Approach: Rule Based System

MS2

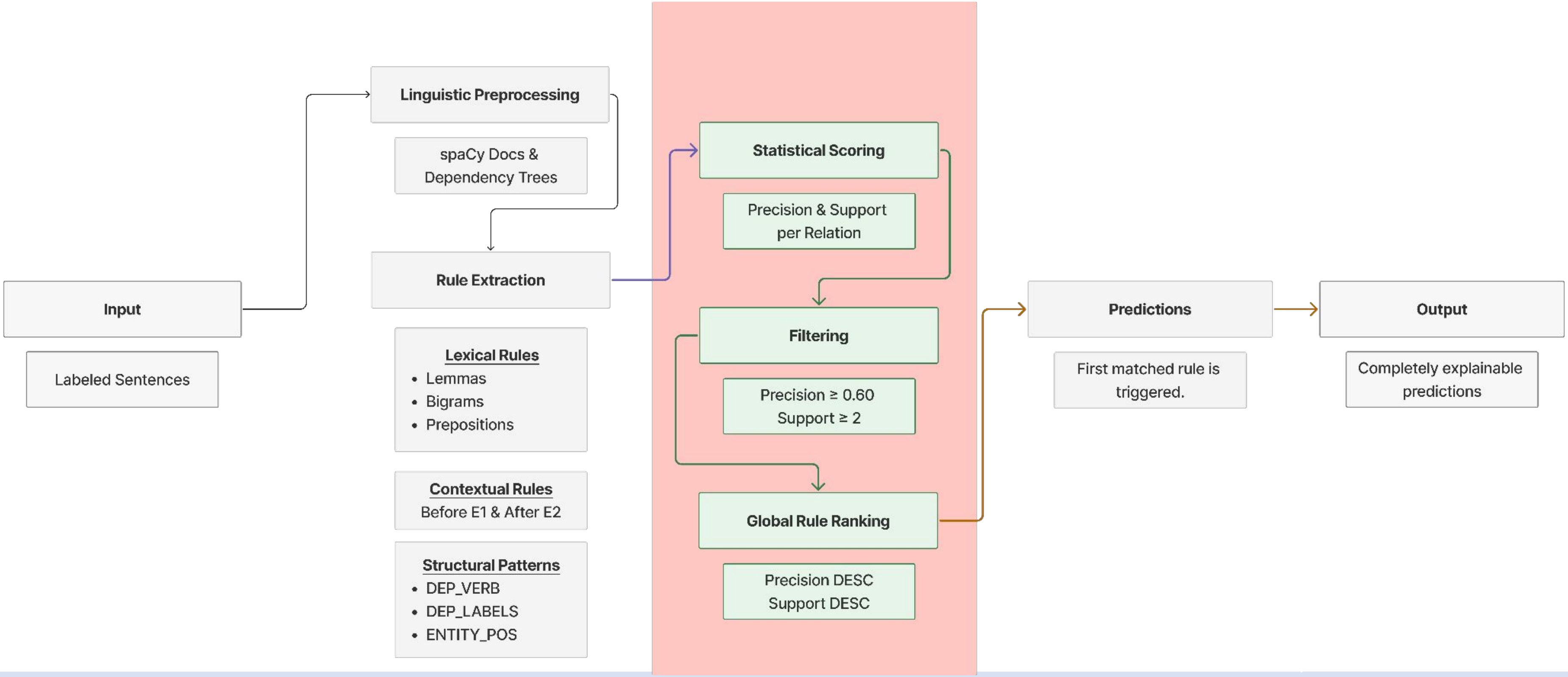
Linguistic Preprocessing & Rule Extraction



Approach: Rule Based System

MS2

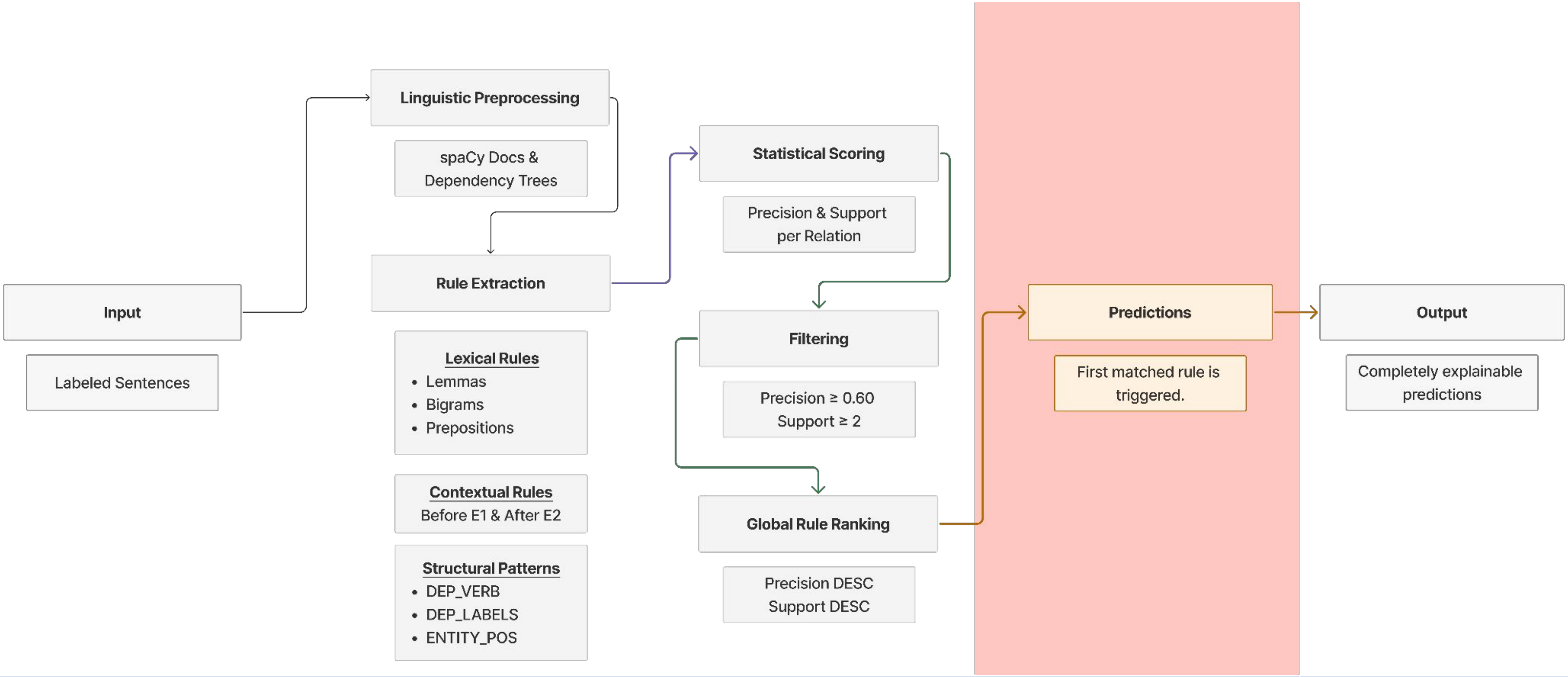
How are the rules chosen?



Approach: Rule Based System

MS2

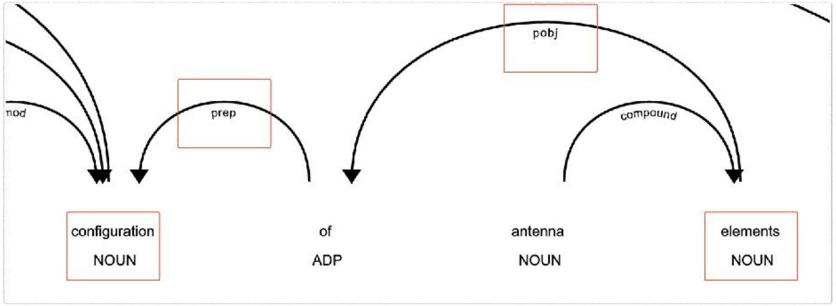
Predictions



Approach: Rule Based System

MS3

SpaCy Dependency Matcher




```
pattern_component_prep_whole = [
    # 1. The COMPONENT (Head)
    {
        "RIGHT_ID": "component",
        "RIGHT_ATTRS": {"POS": "NOUN"}
    },
    # 2. The PREPOSITION
    {
        "LEFT_ID": "component",
        "REL_OP": ">",
        "RIGHT_ID": "prep_word",
        "RIGHT_ATTRS": {
            "DEP": "prep",
            "LOWER": {"IN": ["of", "in", "within", "on", "inside"]}
        }
    },
    # 3. The WHOLE (Target)
    {
        "LEFT_ID": "prep_word",
        "REL_OP": ">",
        "RIGHT_ID": "whole",
        "RIGHT_ATTRS": {"DEP": "pobj", "POS": {"IN": ["NOUN", "PROPN"]}}
    }
]
```

Approach: Rule Based System

MS3

Rule Extraction Enriched

Lexical Rules
- Lemmas
- Bigrams
- Prepositions

Contextual Rules

Before E1 & After E2

Structural Patterns
- DEP_VERB
- DEP_LABELS
- ENTITY_POS

+

SpaCy Rules
- **TRIANGLE** (head/depented operators)
- **BRIDGE** (chained head operators)
- **LINEAR** (no direct depenceny)

e1 ←nsubj- wrote →dobj→ e2

author wrote **book**

e1 ← anchor →prep→ prep_word →pobj→ e2

member of **committee**

e1 … verb … e2

Machine Learning Based Models

Another Approach: Machine Learning Based System

MS2

Overall Results – What We Observed

- Models were evaluated using Accuracy, Precision, Recall, and F1-score
- Linear models achieved the highest overall performance
- Best models reached around 50–55% weighted F1-score
- Complex models failed to surpass linear model results
- This confirms model simplicity worked better for this dataset

SGD Logistic & SGD SVM – Actual Results

- Both models achieved similar accuracy and F1-scores
- Weighted F1-score was approximately 0.50–0.55
- Precision and recall were balanced across classes
- Results indicate good generalization on unseen data
- These models produced the most reliable and stable results

Another Approach: Machine Learning Based System

MS2

Random Forest – Observed Results

- Random Forest achieved lower accuracy than linear models
- F1-score dropped compared to SGD-based models
- High variance with inconsistent class-wise predictions
- Model complexity did not translate into better performance
- Results show Random Forest is not effective for sparse text features

Results: Rule Based System

MS2

Milestone 2 Baseline Results

Train Set Results

Metric	precision	recall	f1-score	support
accuracy			0.58	8000
macro avg	0.71	0.48	0.53	8000
weighted avg	0.69	0.58	0.57	8000

Test Set Results

Metric	precision	recall	f1-score	support
accuracy			0.497	2717
macro avg	0.563	0.402	0.430	2717
weighted avg	0.564	0.497	0.486	2717

Test set performance per relation class (MS2 Baseline)

Relation	Precision	Recall	F1-score	Support
Cause–Effect (e1,e2)	0.837	0.806	0.821	134
Cause–Effect (e2,e1)	0.749	0.722	0.735	194
Component–Whole (e1,e2)	0.353	0.037	0.067	162
Component–Whole (e2,e1)	0.471	0.373	0.416	150
Content–Container (e1,e2)	0.644	0.817	0.720	153
Content–Container (e2,e1)	0.857	0.308	0.453	39
Entity–Destination (e1,e2)	0.722	0.849	0.780	291
Entity–Destination (e2,e1)	0.000	0.000	0.000	1
Entity–Origin (e1,e2)	0.800	0.682	0.737	211
Entity–Origin (e2,e1)	0.600	0.064	0.115	47
Instrument–Agency (e1,e2)	0.500	0.318	0.389	22
Instrument–Agency (e2,e1)	0.526	0.448	0.484	134
Member–Collection (e1,e2)	0.400	0.125	0.190	32
Member–Collection (e2,e1)	0.579	0.109	0.184	201
Message–Topic (e1,e2)	0.648	0.552	0.596	210
Message–Topic (e2,e1)	0.629	0.431	0.512	51
Other	0.210	0.476	0.291	454
Product–Producer (e1,e2)	0.640	0.148	0.241	108
Product–Producer (e2,e1)	0.523	0.366	0.431	123
Accuracy			0.497	2717
Macro Avg	0.563	0.402	0.430	2717
Weighted Avg	0.564	0.497	0.486	2717

Insight 1

MS3

Relation Type	precision	recall
Other	0.210	0.476

Problem:

We have low precision in "*Other*" type.

Cause:

Predicting *"Other"* type by default if there is no match.

Solution:

Extracting rules for "*Other*" as well.

Relation Type	precision	recall
Other	0.31	0.21

Insight 2

MS3

Relation Type	precision	recall
Entity-Destination (e2,e1)	0.000	0.000

Problem:

Relations with reverse directions missed

Cause:

Passive structures could not matched

Solution:

Add passive augmentation for movement verbs

Example:

X “moved to” Y ->

Y “was reached by” Z

Relation Type	precision	recall
Entity-Destination (e2,e1)	1	1

Insight 3

MS3

Problem:

Same syntactic pattern can express different relations

Cause:

Syntax alone does not encode entity meaning

Example:

- *timer (n)* of *device (n)*
- *members (n)* of *committee (n)*
- *book (n)* of *author (n)*

Solution:

Use WordNet and Framenet to capture semantics

Group By Category:

timer **(ARTIFACT)** of *device* **(ARTIFACT)** → **Component–Whole**

members **(PERSON)** of *committee* **(GROUP)** → **Member–Collection**

book **(COMMUNICATION)** of *author* **(PERSON)** → **Product–Producer**

Insight 3

MS3

Problem:

Same syntactic pattern can express different relations

Cause:

Syntax alone does not encode entity meaning

Example:

- *timer (n)* of *device (n)*
- *members (n)* of *committee (n)*
- *book (n)* of *author (n)*

Solution:

Use WordNet and Framenet to capture semantics

Results:

Metrics	Before	After	Change
Accuracy	49.7%	**57.6%**	+7.9%
Macro Avg Precision	56.3%	**61.3%**	+5%
Macro Avg Recall	40.2%	**49.8%**	+9.6%

Conclusions

MS3

- Transformers outperform RB methods (~80% accuracy)
- They capture semantic variation and complex linguistic patterns

At the same time;

- RB systems remain valuable for high-precision settings
- RB methods offer strong interpretability and can be scaled with semantic resources

Questions

Backup: Our Current Results By Category

MS3

Test Set Results (MS2 Baseline + Semantic Patterns)

Relation	Precision	Recall	F1-score	Support
Cause-Effect (e1,e2)	0.858	0.813	0.835	134
Cause-Effect (e2,e1)	0.750	0.820	0.783	194
Component-Whole (e1,e2)	0.711	0.364	0.482	162
Component-Whole (e2,e1)	0.496	0.400	0.443	150
Content-Container (e1,e2)	0.667	0.810	0.732	153
Content-Container (e2,e1)	0.704	0.487	0.576	39
Entity-Destination (e1,e2)	0.732	0.890	0.803	291
Entity-Destination (e2,e1)	0.000	0.000	0.000	1
Entity-Origin (e1,e2)	0.771	0.796	0.783	211
Entity-Origin (e2,e1)	0.800	0.255	0.387	47
Instrument-Agency (e1,e2)	0.636	0.318	0.424	22
Instrument-Agency (e2,e1)	0.509	0.440	0.472	134
Member-Collection (e1,e2)	0.600	0.281	0.383	32
Member-Collection (e2,e1)	0.697	0.537	0.607	201
Message-Topic (e1,e2)	0.638	0.605	0.621	210
Message-Topic (e2,e1)	0.574	0.529	0.551	51
Other	0.251	0.381	0.303	454
Product-Producer (e1,e2)	0.735	0.333	0.459	108
Product-Producer (e2,e1)	0.527	0.398	0.454	123
Accuracy			0.576	2717
Macro Avg	0.613	0.498	0.531	2717
Weighted Avg	0.609	0.576	0.577	2717

Additionally if they want to see they may check the results of ms2 + semantics patterns per
relation test results

If we put all of the results (bert, rb) for general comparison (we don't have to present just say for interested ones you can see them in this slide)

There can be also just macro avgs of precision, f1 recall and also accuracy.

Additionally we have good improvement on **recall** for some specific relations
Such as component whole 30%