Decoding Legal Documents: Natural Language Processing and Text Mining for Information Extraction

Irpan Abdurahman, Xinyang Song, Jincai Zheng,
Lucas Tyler, Tianjun Chen, Aaron Wang

NORTON ROSE FULBRIGHT

Introduction

M&A Specialty Insurance Company *(Issuer)*
123 New York Street, New York, New York 10041 *(Issuer address)*

Home Office: c/o NYNYNY Corporation, 321 York New Street , New York, New York 10041
(Issuer corporate or parent / home office address – is present)

THIS IS A CLAIMS MADE AND REPORTED POLICY. SUBJECT TO THE TERMS AND CONDITIONS OF THIS POLICY, COVERAGE IS LIMITED TO CLAIMS THAT THE NAMED INSURED REPORTS TO THE INSURER DURING THE POLICY PERIOD OR WITHIN THE SIXTY (60) DAY PERIOD IMMEDIATELY FOLLOWING THE EXPIRATION OF THE POLICY PERIOD. THE LIMIT OF LIABILITY TO PAY JUDGMENTS OR SETTLEMENT AMOUNTS SHALL BE REDUCED AND MAY BE EXHAUSTED BY PAYMENT OF DEFENSE COSTS. PLEASE READ THIS POLICY CAREFULLY.

Item 1: Named Insured: M&A Acquisition CO, LLC *(PURCHASER)*
c/o M&A Capital Partners, LLC *(PURCHASER PARENT – if present)*
100 Rice Avenue
Houston, TX, 77002
Attn: General Counsel
Phone: 832-555-5555
E-mail: GC@MAacqco.com *(PURCHASER INFO)*

Item 2: Additional Insureds: Purchaser Indemnified Parties and the current and future Affiliates of the Named Insured and their respective Subsidiaries, partners, officers, directors, employees, agents and representatives.

Item 3: Policy Period for **Fundamental Representations** and the **Pre-Closing Tax Indemnity**:
From: February 28, 2021 To: [●], 2027
At 12:01 A.M. Standard Time at the Named Insured's mailing address stated in Item 1

Policy Period for **Non-Fundamental Representations**:
From: February 28, 2021 To: [●], 2024
At 12:01 A.M. Standard Time at the Named Insured's mailing address stated in Item 1

Item 4: Limit of Liability: $25,000,000 in the aggregate, excess of the Retention

Item 5: Retention: $5,700,000, in the aggregate until the Retention Dropdown Date.
To the extent the then-remaining retention is greater than $3,800,000 on the Retention Dropdown Date, the Retention would drop to $3,800,000, in the aggregate.

Item 6: Premium: $1,025,000

Item 7: Deal Team Members: John Doe and Jane Smith

Item 8: Purchase and Sale Agreement, dated as of DATE, 2021 made and entered into by and between SELLER , STATE & Business Structure, and PURCHASER , STATE & Business Structure *(Seller & Buyer)*

Legal Insurance Policy

Can you identify the "Limit of Liability" value from this document?

How long did it take?

What if you need to do it for 100 documents like this?

Our Solution: Use Named Entity Recognition and Text Parsing to extract key entities from legal documents.

Objectives

Unstructured Legal Documents

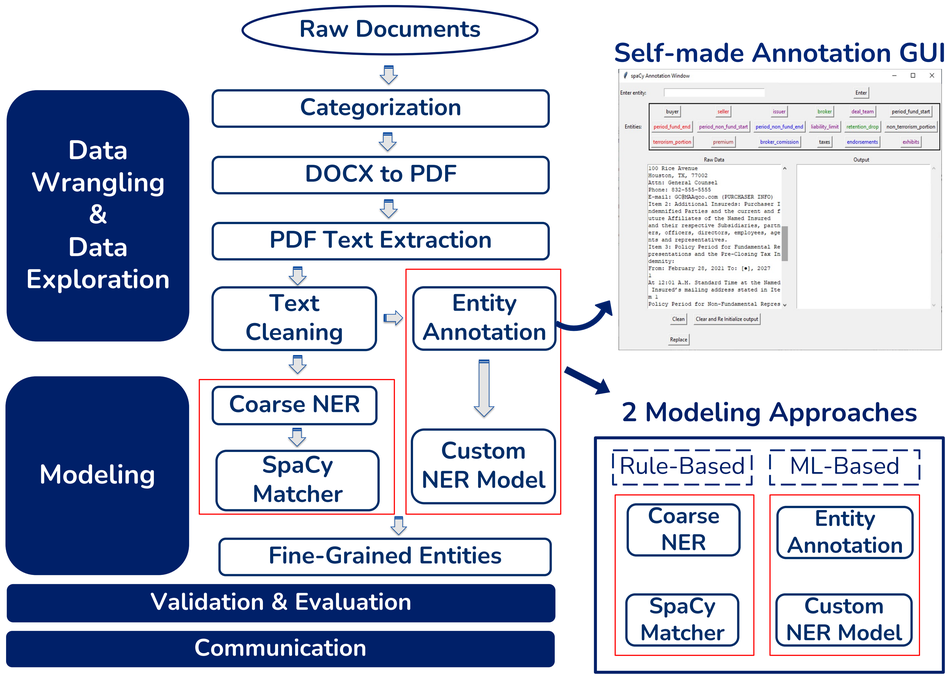
Doc ID	Premium	...	Policy Ends
001	$125,000	...	02/28/2024
002	$800,000	...	06/15/2018
003	$30,000	...	11/30/2024
004	$350,000	...	01/05/2025
005	$700,000	...	12/31/2024

Structured Database

Coarse Entity Type	Fine-Grained Entity Type
ID	doc_id
Organization	policy_purchaser issuer broker deal_team
Date	period_fund_start period_fund_start period_non_fund_start period_non_fund_end
Money	premium liability_limit retention retention_drop

13 Key Entities to Extract from the Legal Documents

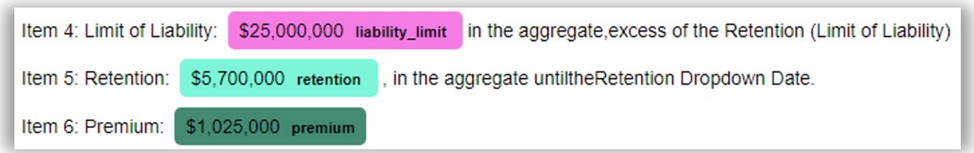
Pipeline



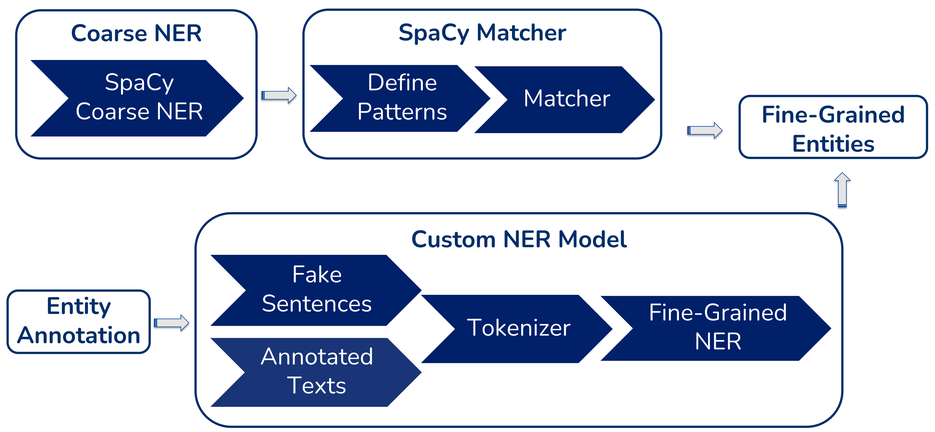
Self-made Annotation GUI

2 Modeling Approaches

Methodology

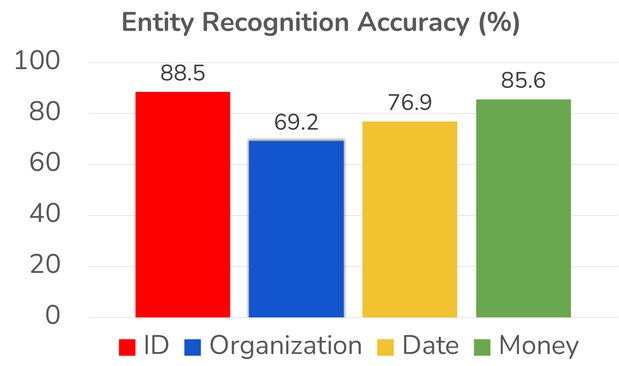


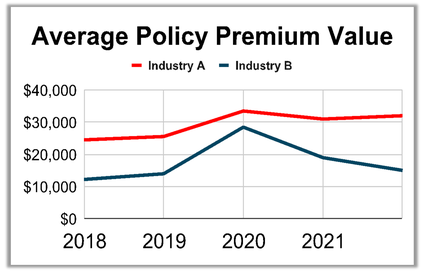
Named Entity Recognition (NER) Example



Rule-based (Top) and ML-based (Bottom) Modeling Methodologies

Results





Example Use-Case with Extracted Entities

Conclusion

Rule-Based	ML-Based
• High accuracy • Requires maintenance	• More use-cases • Need more training data

Future Work

- **Text Extraction**
  1. Automate text extraction module
- **Rule-Based Model**
  1. Use Regex (Time efficiency)
  2. More rules for new entity patterns.
- **ML-Based Model**
  1. More training data (at least 500+ documents).
  2. Use the rule-based model to label training data.

Acknowledgements

Thank you to our sponsors from Norton Rose Fulbright: Zack Barnes, Alper Tuzun, Alexis Collins, and faculty mentor Dr. Arko Barman.