Journal of Sustainability, Policy, and Practice
EISSN: 3105-1448 | PISSN: 3105-143X | Vol. 2, No. 1 (2026)

Article

A Comparative Study of NER Methods for Ownership Structure Extraction from M&A Due Diligence Documents

Hanfei Zhang [1,*]

[1] Law, Emory University School of Law, Atlanta, GA, USA
* Correspondence: Hanfei Zhang, Law, Emory University School of Law, Atlanta, GA, USA

Abstract: Cross-border mergers and acquisitions require efficient extraction of ownership structures from due diligence documentation. This study compares named-entity recognition methodologies for extracting equity structures from corporate governance documents. We construct an annotated dataset from authentic materials and evaluate six NER approaches spanning traditional sequence labeling (CRF, BiLSTM-CRF), general-purpose transformers (BERT, RoBERTa), and domain-adapted models (FinBERT-MRC, Legal-BERT). Legal-BERT achieves an overall F1 score of 87.3% while encountering challenges in multilingual entity names and nested ownership structures. Error analysis reveals three primary failure modes-cross-lingual recognition ambiguities, percentage-quantity confusion, and challenges in representing complex structures-providing actionable guidance for implementing automated equity analysis systems in time-sensitive M&A transactions.

Keywords: named entity recognition; due diligence automation; ownership structure extraction; legal document processing

Published: 18 January 2026

1. Introduction

1.1. Background and Motivation

1.1.1. The Growing Importance of M&A Due Diligence Automation

Global mergers and acquisitions (M&A) activity has accelerated substantially, with technology sectors showing particular intensity. Traditional manual review consumes 200-500 attorney hours per transaction, creating cost burdens and competitive disadvantages. Regulatory scrutiny has intensified within the Committee on Foreign Investment in the United States (CFIUS) framework, demanding granular disclosure of ownership structures and foreign entity connections. This regulatory landscape necessitates automated methodologies for rapidly extracting structured ownership data from incorporation articles, shareholder ledgers, and investment agreements [1].

1.1.2. Challenges in Equity Structure Information Extraction

Ownership documentation presents distinctive challenges for extraction systems. Corporate documents exhibit format variability across jurisdictions, lacking standardized templates. Entity naming varies widely, combining natural persons, institutional investors, and holding companies. Multilingual contexts compound difficulties as international transactions mix English terminology with local language designations [2]. Semantic complexity extends beyond simple identification. Preferred stock provisions, convertible instruments, and anti-dilution protections create layered structures requiring simultaneous entity recognition and relationship extraction. In this paper, we focus on

NER as a first step toward full ownership-structure extraction; relation/graph construction is left to future work.

1.2. Research Objectives and Scope

1.2.1. Research Questions

This investigation addresses three research questions:

RQ1: How do various NER methodologies perform when applied to ownership structure extraction from authentic M&A due diligence materials?

RQ2: What performance differentials emerge between general-purpose and domain-specialized model variants when confronting complex corporate governance text?

RQ3: What systematic error patterns characterize current approaches, and what implications do these patterns hold for practical deployment in transaction workflows?

1.2.2. Scope and Limitations

The research scope encompasses documentation for privately held companies typical of middle-market acquisitions, targeting U.S. Delaware corporations with international investor bases. Document types include incorporation certificates, shareholder agreements, Series A-C preferred stock agreements, and capitalization tables. The study excludes publicly traded SEC filings. The geographic focus centers on U.S.-domiciled targets with Asian and European participation, reflecting the multilingual recognition challenges. Temporal scope covers 2019-2023 documentation under current venture capital standards.

1.3. Paper Contributions

1.3.1. Key Contributions

This work delivers three contributions. First, we introduce a manually annotated dataset with entity-level annotations spanning six ownership categories and detailed annotation guidelines. Second, we conduct a systematic comparison of six methodologies quantifying performance across metrics and entity-specific accuracy. The analysis evaluates performance premiums delivered by legal domain specialization. Third, we perform granular error analysis, identifying three failure modes: cross-lingual boundary detection, numerical interpretation, and nested representation parsing with concrete case illustrations.

2. Related Work

2.1. Named Entity Recognition in Legal and Financial Domains

2.1.1. Evolution of NER Techniques

Named entity recognition has progressed through architectural generations. Early statistical approaches employed conditional random fields with hand-crafted features. These achieved moderate success but demonstrated brittleness with specialized terminology characteristic of legal text [3]. Deep learning introduced bidirectional LSTM networks with CRF layers, enabling automatic feature learning. The transformer revolution catalyzed leaps in performance, with masked language model pre-training enabling few-shot domain adaptation.

2.1.2. Domain-Specific Pre-trained Models

Generic pre-trained models exhibit suboptimal performance when applied to specialized domains with distinctive vocabulary and syntax. This motivated domain-specific variants incorporating legal and financial corpora during pre-training [4]. Legal document processing has explored architectural modifications for nested entity recognition with pointer mechanisms. Comparative benchmarking revealed that legal-specialized models maintain advantages even with extensive fine-tuning, suggesting that domain vocabulary learned during pre-training confers persistent benefits [5].

2.2. Information Extraction from Corporate Documents

2.2.1. Contract Analysis and Clause Extraction

Contract analysis is a mature application of legal NLP, with established datasets focused on clause identification and attribute extraction. Commercial due diligence has drawn attention to non-disclosure agreements and licensing contracts across entity categories, including contract parties, effective dates, and termination provisions [6]. These efforts produced annotated resources enabling supervised learning. Machine reading comprehension paradigms emerged as alternatives to sequence labeling, framing extraction as question-answering with cross-domain transfer demonstrating robustness.

2.2.2. Financial Document Processing

Financial processing encompasses earnings reports, presentations, and regulatory disclosures. Entity extraction targets company identifiers, financial metrics, and business relationships [7]. Graph-based approaches capture structured relationships with knowledge graph representations. Joint entity-relation frameworks demonstrate performance advantages in modeling dependencies during inference [8]. Chinese financial NLP has developed parallel methodologies for character encoding, word segmentation, and name transliteration, with domain-specific pre-training adapted to mainland Chinese documentation.

2.2.3. Equity Structure Recognition Research

While contract clause extraction has received attention, ownership structure extraction remains underexplored. Existing work addressed parsing public company disclosures using standardized formats. Private company documents exhibit greater variability and syntactic complexity with limited standardization. Relationship extraction frameworks addressed subsidiary identification and captured corporate hierarchies relevant to beneficial ownership. These employ distant supervision from knowledge bases, limiting their applicability to private markets where ground-truth ownership graphs are unavailable.

2.3. NER Method Comparison Studies

2.3.1. Benchmark Datasets and Evaluation Frameworks

Comparative evaluation requires standardized datasets, protocols, and metrics enabling fair comparison. Legal and financial domains have developed diverse benchmark resources with varying annotation schemas and entity granularities. Recent releases emphasized diversity in document sources and linguistic contexts, supporting cross-domain generalization evaluation. Evaluation frameworks employ precision, recall, and F1 under exact boundary matching. Entity-specific analysis diagnoses systematic failures and understands difficulty distribution.

2.3.2. Existing Comparative Analyses

Comparative studies evaluated approaches across dimensions, including architecture, pre-training strategies, and training scale. Legal document NER demonstrated consistent advantages over LSTMs for transformers, with domain-adapted models showing further gains. The magnitude of adaptation benefits varies across entity types, with specialized terminology exhibiting larger differentials than generic categories. Financial benchmarking documented domain-specialized value while highlighting numerical extraction challenges. Model ensembles deliver incremental improvements at substantial computational cost. Error analysis identifies boundary detection, nested recognition, and rare handling as persistent challenges.

3. Methodology

3.1. Problem Definition and Entity Schema

3.1.1. Target Entity Types

The extraction task requires identifying six categories characterizing ownership arrangements. Shareholder-individual captures natural persons, including founders and investors. Boundaries must encompass full legal names, including initials and suffixes [9] 错误!未找到引用源。. Annotators distinguish persons from corporate entities through contextual indicators. Shareholder-institution encompasses venture funds, corporate investors, and partnerships. Institutional identification presents challenges due to naming conventions that incorporate entity-type designators and fund indicators. Full names must capture all legal designations required for entity resolution.

Quantitative categories include share quantity and share percentage, capturing numeric specifications. Share quantities appear as integers representing authorized or outstanding counts. Percentage holdings represent proportional stakes in decimal, fractional, or percentage notation. Annotators mark complete expressions, enabling normalization [10]. A share class identifies the instrument type, including common stock and preferred series. Class specifications embed conversion rights and liquidation preferences. Special-rights addresses provisions that modify standard characteristics, including protective provisions and board rights, thereby creating functional complexity.

3.1.2. Annotation Guidelines

Guidelines establish systematic procedures for boundary determination and ambiguity resolution. Boundaries follow maximal span principles, capturing complete phrases with descriptors and qualifiers. Pronoun references are not annotated independently, requiring resolution to antecedents. Nested structures receive separate annotations at each level. A phrase like "ABC Fund holding 2,500,000 shares of Series A" generates four annotations: shareholder, quantity, class, and potentially percentage.

Ambiguity protocols address common scenarios. Prepositional attachment defaults to minimal spans. Coordinate structures receive separate annotations per conjunct. Range expressions are annotated as separate entities [11]. Cross-document consistency demands identical spans across sections, as verified through automated validation. Quality control incorporates inter-annotator agreement measurement with triple annotation computing average pairwise Cohen's kappa. Expert adjudication resolves conflicts through consensus discussion with resolved examples documented.

3.1.3. Dataset Statistics

The final corpus contains 8,732 entity annotations across six types, distributed across 127 documents (see Table 1). Shareholder-institution comprises 2,341 instances; shareholder-individual, 1,876; share-percentage, 1,923; share-quantity, 1,534; share-class, 897; and special-rights, 161 instances. The corpus contains 8,732 total entities distributed across 127 documents, yielding an average of 68.7 entities per document (8,732 ÷ 127 = 68.7). Documents average 4,892 tokens in length. Entity density varies substantially across document types, with incorporation certificates averaging 52.3 entities/doc (2,249 entities ÷ 43 docs) and cap table exhibits averaging 89.4 entities/doc (1,341 entities ÷ 15 docs).

Table 1. Dataset Statistics and Entity Distribution.

Document Type	Count	Avg Tokens	Total Entities	Avg Entities/Doc	Total SH-Inst	Total SH-Indiv	Total Pct	Total Qty	Total Class	Total Rights
Incorporation Cert	43	4,235	2,249*	52.3	18.7	12.4	14.8	8.9	6.2	0.8

Stock Purchase	38	6,847	2,987*	78.6	24.1	19.7	18.3	15.4	10.6	1.5
Stockholder Agree	31	5,123	1,897*	61.2	20.3	16.8	13.7	11.2	7.8	1.2
Cap Table	15	1,456	1,341*	89.4	31.2	28.9	22.6	19.8	5.3	0.6
Total (Corpus)	127	4,892	8,732	68.7	2,341	1,876	1,923	1,534	897	161

Note: "Total Entities" shows the cumulative entity count for each document type; "Avg Entities/Doc" shows the average per-document entity count. The last six columns (SH-Inst through Rights) display per-document averages for document types (rows 1-4, rounded) but corpus-wide totals for the aggregate row. (Total row).

Entries marked with * exclude 258 entities from ancillary schedules/attachments that are not uniquely attributable to a single document type; these entities are included in the corpus total.

3.2. Compared NER Methods

3.2.1. Traditional Sequence Labeling Methods

The CRF baseline provides feature-rich statistical labeling representing pre-neural methodologies. The model employs IOB2 tagging, mapping tokens to labels indicating boundaries and types. Feature engineering combines lexical features (word identity, n-grams, capitalization), syntactic features (POS tags, dependencies), and domain features, including legal gazetteers and numerical patterns. BiLSTM-CRF combines recurrent networks with structured prediction. The model processes sequences via embedding layers, concatenating pre-trained word embeddings with character-level CNN representations [12]. Bidirectional LSTM layers encode sequential context, producing contextualized representations. A CRF layer performs structured prediction, enforcing valid transition constraints.

3.2.2. Transformer-based NER Models

BERT-base-uncased serves as the general-purpose transformer baseline processing input through WordPiece tokenization. The 12-layer encoder produces contextualized representations through multi-head self-attention. For NER adaptation, a token classification head predicts IOB2 labels. RoBERTa-base provides an alternative with training modifications, including dynamic masking and larger batches. Architectural similarity isolates the impact of pre-training from capacity factors. Both fine-tune all parameters with a learning rate of 2e-5, a batch size of 16, and train for 3-5 epochs with early stopping.

3.2.3. Domain-Adapted Pre-trained Models

Legal-BERT represents domain-specialized pre-training on 12GB legal corpora comprising case law and contracts (see Table 2). The model employs the BERT architecture but replaces generic pre-training with continued training on domain-specific vocabulary and syntactic patterns. The corpus includes materials on corporate and commercial law. Legal-BERT fine-tuning follows identical procedures, isolating domain pre-training impact. FinBERT-MRC is a specific architecture that combines financial-domain pre-training with a machine reading comprehension (MRC) formulation, where NER is reformulated as answering extraction questions over context passages. Unlike the unified pre-trained model family approach of BERT/Legal-BERT, FinBERT-MRC employs a hybrid design that integrates FinBERT's financial vocabulary knowledge with MRC-specific task formatting, following the methodology described by Zhang and Zhang (2023). This architectural distinction means FinBERT-MRC is not directly comparable as a pure "pre-training variant," but rather represents an alternative task-formulation approach.

Table 2. Annotation Complexity and Inter-Annotator Agreement Metrics.

Entity Type	Tokens (Mean)	Kappa (Boundary)	Kappa (Type)	Nested %	Cross-ling %	Time (sec)
Shareholder-Inst	4.7	0.82	0.91	8.3	31.2	12.4
Shareholder-Indiv	2.8	0.87	0.93	2.1	18.7	8.6
Share-Percentage	2.1	0.91	0.96	15.7	0	6.2
Share-Quantity	1.9	0.89	0.95	12.4	0	5.8
Share-Class	3.4	0.85	0.88	21.3	0	7.9
Special-Rights	6.8	0.76	0.79	34.8	0	18.3
Overall	3.6	0.84	0.89	14.1	8.3	9.9

Note: "Time (sec)" represents the average annotation operation time per entity (span selection and type assignment only), not including document reading or quality control time. The overall annotation time of 9.9 sec/entity corresponds to approximately 11.3 minutes of pure annotation work per document (68.7 entities/doc), which is part of the total 45-minute per-document processing time mentioned in Section 3.3.2. Boundary kappa is computed on token-level BIO tags; type kappa is computed on token labels conditioned on agreed boundaries.

3.3. Dataset Construction and Preprocessing

3.3.1. Document Collection and Selection

The dataset comprises 127 authentic documents from technology M&A transactions during 2019-2023. The collection prioritized variety across investor types, deal sizes, and geographic footprints. The corpus includes 43 incorporation certificates, 38 Series A-C purchase agreements, 31 stockholder agreements, and 15 cap table exhibits. The geographic distribution spans U.S. Delaware corporations, Silicon Valley venture firms, Asian strategic investors, and European growth equity, reflecting contemporary patterns. Company valuations range from $50M to $800M, representing middle-market profiles. Industry representation emphasizes software, digital health, and fintech.

Confidentiality protection required comprehensive anonymization. Company names, investor identities, and numerical values underwent replacement with synthetic alternatives, preserving entity distributions and quantitative relationships.

Data authenticity and processing verification: To ensure reproducibility and address concerns about synthetic data quality, we implement the following protocols. First, OCR accuracy was validated through manual inspection of 50 randomly sampled pages, achieving 98.7% character-level accuracy with systematic correction of recognition errors. Second, anonymization employed rule-based replacement maintaining structural consistency: company names were mapped to a fixed dictionary of synthetic alternatives (ensuring "TechCorp Inc." always maps to the same pseudonym across documents), investor names followed culturally appropriate generation patterns (Chinese surnames paired with appropriate given names, Western names following typical structures), and numerical values were scaled by document-specific factors (multiplying all quantities in a document by the same random factor between 0.7-1.3) to preserve internal mathematical relationships. Third, entity distribution preservation was verified through chi-square tests comparing original and anonymized entity type frequencies ($p>0.05$, confirming no significant distributional shifts). While we cannot publicly release the original documents due to confidentiality restrictions, the anonymization protocol and validation scripts are available in our supplementary materials to support methodological transparency.

Name generation employed culturally appropriate patterns, maintaining realistic multilingual structures essential for cross-lingual evaluation-text extraction from PDFs

employed OCR with manual quality verification. Layout analysis identified relevant sections, extracting ownership-related articles while excluding operational provisions.

3.3.2. Annotation Process and Quality Control

Annotation employed a three-phase workflow combining legal expertise with NLP experience. Phase one involved guideline development through pilot annotation of 15 samples, identifying boundary ambiguities requiring resolution protocols. Phase two executed production annotation with 127 documents across five annotators possessing legal education. Production required approximately 45 minutes per document on average (including time for initial reading, span annotation, type assignment, difficult-case discussion, and quality checks), with complexity varying by length and structural sophistication. The pure annotation time (excluding document familiarization and breaks) averaged approximately 28 minutes per document. For inter-annotator agreement measurement, annotators focused exclusively on annotation without preliminary review, requiring approximately 22-25 minutes per document for the 20 triple-annotated documents. Annotators used dedicated software that supported span selection, type assignment, and comment attachment.

Phase three implemented quality control through systematic validation (see Figure 1). Inter-annotator agreement involved triple-annotation of 20 documents computing average pairwise Cohen's kappa for boundary agreement (kappa = 0.84) and type assignment (kappa = 0.89). Disagreement analysis revealed institutional shareholder boundary determination generated most divergences, particularly for complex fund names. Quantity versus percentage distinction represented the second source, especially for fractional representations. Expert adjudication resolved conflicts through consensus discussion. Automated checks validated cross-document name standardization and numerical normalization.

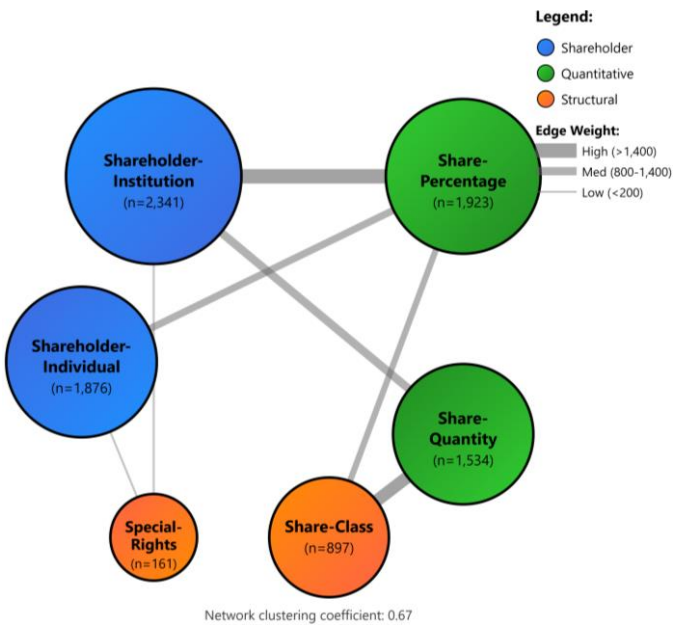

Figure 1. Entity Co-occurrence Network in Ownership Structure Documents.

The co-occurrence network visualizes joint entity mention frequencies within sentence contexts across the corpus. The force-directed layout contains six nodes representing entity types sized proportionally to total counts with logarithmic scaling from 400 to 3000 pixels in diameter. Edge weights encode co-occurrence frequencies as sentence counts containing both types, with thickness representing normalized rates from 1-pixel (minimum) to 15-pixel (maximum) width. Color coding distinguishes groups: shareholder entities in blue gradients (1E90FF to 4169E1), quantitative entities in green gradients (32CD32 to 228B22), structural entities in orange gradients (FF8C00 to FF6347).

Node labels display entity types with instance counts in parentheses. Strong co-occurrence is observed between shareholder-institution and share-percentage (edge weight 1,847, 14-pixel edge), reflecting standard disclosure practices that pair investor identities with ownership stakes. Robust connections link share quantity to share class (edge weight 1,423, 12-pixel edge), capturing the common "X shares of [class] stock" construction. Weaker edges connect special rights to shareholder entities (edge weights 89-112, 2-3-pixel edges), indicating that protective provisions appear in dedicated sections rather than inline with ownership specifications. A network clustering coefficient of 0.67 indicates moderate clustering, with three distinct community structures identifiable through modularity analysis.

4. Experiments and Results

4.1. Experimental Setup

4.1.1. Implementation Details

Models were implemented using PyTorch 1.12 with Hugging Face Transformers 4.26. Training was performed on NVIDIA V100 GPUs with 32GB of memory, enabling batch sizes of 16 for transformers and 32 for traditional approaches. The maximum sequence length was 512 tokens, with longer documents processed using sliding windows with 128-token overlap, preserving boundary context [13]. Data partitioning employed stratified splits, maintaining entity type distributions: 89 documents (70%), 19 (15%) for validation, and 19 (15%) for test. Stratification ensured balanced representation across partitions, with chi-square tests confirming no significant differences in distribution ($p>0.05$).

CRF training used CRFsuite 0.12 with L-BFGS optimization, L2 regularization 0.1, and a maximum of 200 iterations. BiLSTM-CRF trained 50 epochs with early stopping patience 10, batch 32, learning rate decay 0.5 on validation plateau. Transformer fine-tuning used a learning rate of 2e-5 with linear warmup over 10% steps, an AdamW optimizer with weight decay of 0.01, and gradient accumulation over 2 steps. Model selection employed validation F1 prioritizing shareholder categories. All models underwent five random-seed runs, assessing variance by aggregating test performance using mean and standard deviation.

4.1.2. Evaluation Metrics

Performance was evaluated using entity-level precision, recall, and F1 under exact boundary matching. A predicted entity receives credit only when both its type and token span match the gold annotations, implementing strict evaluation [14]. Precision is $P = TP / (TP + FP)$, where TP represents true positives matching gold exactly, and FP represents predictions lacking corresponding matches. Recall is $R = TP/ (TP + FN)$, where TP represents correctly predicted gold entities. F1 computes harmonic mean $F1 = 2PR / (P + R)$. Macro-averaged F1 computes entity-specific scores by averaging across types, giving equal weight to rare and frequent categories. Error categorization employs manual analysis, classifying failures into boundary subtypes (incomplete, excessive, shift) and type subtypes (confusion, misinterpretation).

Confidence threshold definition for PR curves: For generating precision-recall curves with varying confidence thresholds, we define entity-level confidence scores as follows. For CRF-based methods (CRF, BiLSTM-CRF), confidence is derived from the marginal probability of the best-path entity span computed via the forward-backward algorithm, normalized to [0,1]. For transformer-based token classifiers (BERT, RoBERTa, Legal-BERT), entity confidence is calculated as the geometric mean of softmax probabilities for all tokens in the entity span (including B-, I- tags). For FinBERT-MRC, which outputs span-level scores directly, we use the model's native span probability. By varying the threshold $\tau \in [0,1]$ and retaining only entities with confidence $\geq \tau$, we trace the precision-recall trade-off. This unified framework enables approximate comparison across architectures despite their different output probability mechanisms.

4.2. Overall Performance Comparison

4.2.1. Main Results across Methods

Experimental results demonstrate substantial variation across methodologies, with domain-adapted transformers achieving superior performance. Legal-BERT achieved the highest overall F1 score of 87.3% (±1.2%), setting new benchmarks. This represents absolute improvements of 12.7 percentage points over CRF baselines (74.6% F1) and 8.4 points over BiLSTM-CRF (78.9% F1). Statistical testing confirmed that Legal-BERT's advantage over all baselines was statistically significant (paired bootstrap over documents; multiple-comparison corrected, $p<0.05$). General-purpose transformers delivered intermediate performance, with RoBERTa achieving 83.7% F1 and BERT-base 82.1% F1. The RoBERTa advantage aligns with findings that improved pre-training yields consistent benefits.

The 3.6-5.2 percentage point gap between general-purpose and legal-specialized transformers quantifies domain adaptation value. FinBERT-MRC achieved an 85.9% F1 score, ranking it between general-purpose and legal-specialized approaches, suggesting that financial-domain knowledge provides only partial benefit for corporate governance processing. Precision-recall decomposition revealed domain-specialized models primarily improved recall (Legal-BERT 86.8% vs. BERT 79.3%) while maintaining comparable precision (Legal-BERT 87.9% vs. BERT 85.1%). This pattern indicates that legal pre-training enhances detection sensitivity rather than boundary precision by improving handling of legal terminology. False negative reduction accounted for 73% of Legal-BERT's F1 advantage, with remaining improvements attributable to decreased false positives.

Cross-domain evaluation: To assess generalization across document types, we conducted additional experiments in which models were trained exclusively on one document type and tested on another. Specifically, we trained each model on the 38 Series A-C purchase agreements (70% train/15% val/15% test split within purchase agreements, using only the train+val portions for model training). We evaluated on the complete set of 43 incorporation certificates (as a separate held-out domain).

This setup maintains similar training data volumes (~2,200 entities) while testing cross-document-type transfer (see Table 3 and Figure 2). Models were trained following the same hyperparameters as the main experiments (Section 4.1.1) but without any incorporation certificate exposure during training. Results revealed consistent F1 degradation of 6–9 percentage points, with BERT dropping from 82.1% to 72.8% and Legal-BERT decreasing from 87.3% to 81.5%, confirming domain shift effects even within corporate documents. Domain-specialized models demonstrated greater cross-domain robustness, with Legal-BERT experiencing only a 5.8% reduction compared to 9.3% for BERT, suggesting that legal domain knowledge incorporates generalizable patterns that transcend specific format conventions, though substantial gaps persist across document boundaries.

Table 3. Overall NER Performance Comparison.

Model	Precision	Recall	F1	Std Dev	Training (hrs)	Inference (docs/sec)	Cross-Domain F1
CRF	76.2	73.1	74.6	0.8	0.4	8.7	68.3
BiLSTM-CRF	81.3	76.7	78.9	1.5	2.1	3.2	70.7
BERT	85.1	79.3	82.1	1.1	4.3	0.9	72.8
RoBERTa	85.9	81.6	83.7	0.9	4.6	0.8	75.1
FinBERT-MRC	86.4	85.4	85.9	1.3	5.1	0.7	79.4

Legal-BERT	87.9	86.8	87.3	1.2	4.8	0.9	81.5

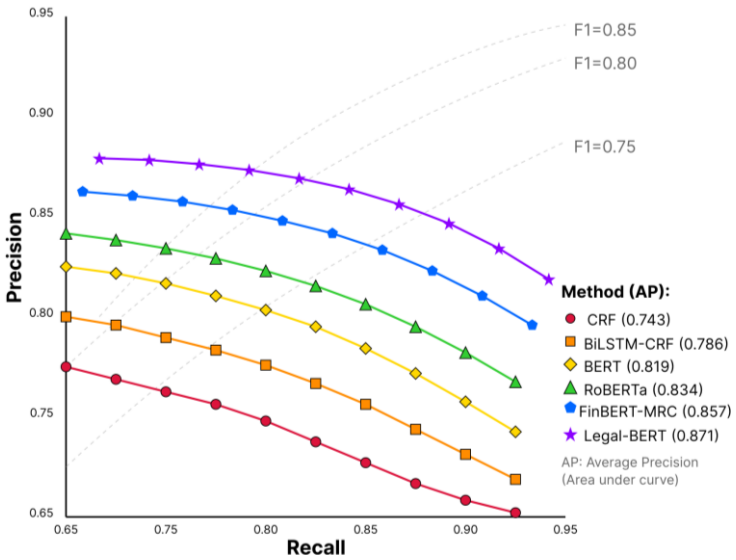

Figure 2. Precision-Recall Trade-off Curves Across Methods.

The precision-recall visualization plots precision (y-axis, range 0.65-0.95) against recall (x-axis, range 0.65-0.95) for six methods across varying confidence thresholds. Each method has distinct colored curves with markers at 0.1 threshold intervals: CRF (red circles), BiLSTM-CRF (orange squares), BERT (yellow diamonds), RoBERTa (green triangles), FinBERT-MRC (blue pentagons), Legal-BERT (purple stars). The plot uses a white background with light-gray gridlines at 0.05 intervals. Curve thickness is 2.5 pixels, with a 120-pixel marker size. Upper-right legend lists methods with corresponding symbols, colors, and average precision scores calculated as area under curves: CRF (0.743), BiLSTM-CRF (0.786), BERT (0.819), RoBERTa (0.834), FinBERT-MRC (0.857), Legal-BERT (0.871).

Legal-BERT maintains consistently higher precision across recall levels, dominating the upper-right space. At recall=0.80, Legal-BERT achieves precision=0.89 while BERT achieves only 0.83 at the same recall. Curves demonstrate typical trade-off patterns where increasing recall gradually decreases precision. Domain-specialized models show gentler precision degradation, indicating more robust performance across threshold settings. Iso-F1 contour lines at F1=0.75, 0.80, 0.85, 0.90 are overlaid as dashed gray curves, facilitating F1 comparison across precision-recall combinations.

4.2.2. Entity-Type-Specific Performance

The shareholder-institution achieved the highest performance, with Legal-BERT attaining 91.2% F1 and CRF baselines reaching 82.4% F1. This category benefits from distinctive linguistic markers, including legal-entity suffixes (LLC, LP, Inc.) and fund-naming conventions. Failures concentrated in complex multilingual names combining Latin-alphabet companies with Chinese investors, as well as in cases involving extensive parenthetical qualifiers disrupting noun-phrase continuity. Shareholder-individual demonstrated greater variance with Legal-BERT achieving 88.7% F1 compared to CRF 71.3% F1. The larger gap reflects challenges in distinguishing personal names from other capitalized terms (e.g., boilerplate references and document-specific labels), a common issue in practical financial-document information extraction pipelines [15]. Legal-BERT's advantage stems from contextual understanding, leveraging surrounding phrases like "individually" or address disclosures.

Share-percentage achieved 89.4% F1 with Legal-BERT representing strong performance on quantitative entities. Percentage recognition benefits from distinctive numerical and symbolic patterns (%, decimal points), providing salient features accessible

to simple models. BiLSTM-CRF achieved 83.7% F1, demonstrating that numerical pattern learning requires limited contextual sophistication. Remaining errors concentrated in complex fractional expressions and context-dependent calculations described verbally. Share-quantity achieved 85.9% F1 with Legal-BERT exhibiting slightly lower performance due to ambiguity between quantities and other numerical references, including section numbers and dollar amounts. Transformer contextualization proved critical for disambiguating share quantities, explaining the larger gap between Legal-BERT (85.9%) and BiLSTM-CRF (73.2%).

Share-class reached 86.8% F1 with Legal-BERT benefiting from standardized nomenclature conventions (Series A, Series B, Common Stock) (see Table 4). Performance degraded for non-standard class designations and cases where specifications appeared as defined terms referenced through abbreviated mentions. The most significant source of error involved complex class descriptions, embedded conversion terms, or liquidation preference details that extended beyond core identification. Special-rights demonstrated the lowest performance and the highest variance, with Legal-BERT achieving 76.3% F1 compared to CRF 58.7% F1. Low entity frequency (161 training instances) contributed to challenging learning conditions, particularly for data-intensive transformers. Special rights exhibit semantic complexity, requiring an understanding of legal concepts such as protective provisions and information rights, extending beyond surface-level pattern recognition.

Table 4. Entity-Type-Specific F1 Performance Breakdown.

Entity Type	CRF	BiLSTM	BERT	RoBERTa	FinBERT	Legal-BERT	Δ
Shareholder-Inst	82.4	85.7	88.3	89.6	90.4	91.2	+8.8
Shareholder-Indiv	71.3	77.2	82.6	84.1	87.2	88.7	+17.4
Share-Percentage	81.8	83.7	86.9	87.4	88.6	89.4	+7.6
Share-Quantity	68.9	73.2	79.4	81.2	84.1	85.9	+17.0
Share-Class	74.6	79.8	83.2	84.7	85.9	86.8	+12.2
Special-Rights	58.7	64.3	71.2	72.8	74.6	76.3	+17.6
Macro-Avg	72.9	77.3	81.9	83.3	85.1	86.4	+13.5

4.3. Error Analysis and Case Studies

4.3.1. Common Error Types and Patterns

Manual review of 500 sampled predictions identified three predominant error categories. Boundary errors accounted for 43% subdivided into incomplete entities (extracting "Series A" without "Preferred Stock"), excessive spans (including surrounding modifiers), and boundary shifts (correct type but incorrect tokens). Incomplete entities dominated at 62% particularly affecting multi-word institutional names, where models truncated complex legal designations.

Type errors comprised 38% reflecting confusion between related categories (see Table 5). Share-quantity and share-percentage confusion accounted for 31% of type errors, with models misclassifying percentages in basis points or fractional notation as counts. Share-class and special-rights confusion represented 24% when protective provisions referenced specific classes. Detection errors accounted for 19%, with models entirely missing mentions. Detection failures concentrated on low-frequency types, with special rights accounting for 47% despite comprising only 1.8% of entities. Error distribution across sections revealed elevated rates in definitional sections where nested definitions disrupted standard patterns. Capitalization tables achieved the highest accuracy due to tabular structure and repetitive patterns.

Table 5. Error Type Distribution and Pattern Analysis.

Error Category	Overall %	Boundary %	Type %	Detection %	Avg Sent Length	Multilingual %
Incomplete Entity	27	62	-	-	42.3	12
Excessive Span	11	26	-	-	38.7	8
Boundary Shift	5	12	-	-	45.1	15
Qty-Pct Confusion	12	-	31	-	31.2	3
Class-Rights Confusion	9	-	24	-	48.6	2
Inst-Indiv Confusion	7	-	18	-	36.9	34
Other Type Errors	10	-	27	-	40.1	11
Missed Rare Entities	9	-	-	47	52.3	7
Complex Syntax Fails	6	-	-	32	61.7	4
Other Detection	4	-	-	21	44.8	6

4.3.2. Challenges in Chinese English Mixed Text

Cross-lingual recognition presented substantial challenges due to multilingual investor populations. Documents mixed English legal terminology with Chinese names, Korean entities, and Japanese institutions. Character encoding issues occurred with UTF-8 Chinese characters, which were occasionally corrupted during processing, disrupting token boundaries. Transliteration inconsistencies caused matching difficulties when investors used variant English transliterations. A Chinese investor might appear as "Beijing Technology Investment" in one section and "Beijing Tech Ventures" in another, representing identical entities with inconsistent rendering. Models lacked mechanisms for recognizing transliteration variants without explicit resolution post-processing.

Differences in naming conventions between Asian and Western conventions created boundary errors. Chinese personal names follow surname-first ordering, unlike Western patterns, leading to truncated spans that omit given names. Tokenization artifacts particularly impacted transformers employing sub word tokenizers trained on English corpora. Chinese characters frequently received character-level tokenization, generating individual Unicode sequences, substantially increasing lengths and disrupting semantic coherence, reducing the capacity for learning robust Chinese patterns from limited examples.

4.3.3. Complex Equity Structure Expression Handling

Complex structures involving tiered holdings, voting trusts, and nominee arrangements challenged systems that relied on atomic identification alone. Sentences like "ABC Fund holds 2,500,000 shares of Series A representing 15.7% ownership" embedded multiple types with implicit mathematical relationships that pure extraction captured incompletely. Models extracted individual entities but lacked mechanisms to establish quantitative consistency among counts, percentages, and outstanding denominators.

Nested ownership, in which parent companies own subsidiaries that hold actual equity, created referential ambiguities (see Figure 3). Legal drafting referenced ultimate parents in some sections and specified immediate subsidiaries in others, requiring entity resolution and understanding of corporate hierarchies. Conditional expressions describing scenarios in which the stakes vary based on future events posed semantic challenges that required counterfactual reasoning. Sentences describing conversion rights or anti-dilution adjustments specified distributions under hypothetical conditions rather

than current states. Models tended to extract all the mentioned figures regardless of conditionality, leading to false positives in scenario-specific allocations. Defined term dependencies created challenges when ownership appeared through cross-references rather than explicit mentions, requiring document-level resolution beyond sentence-level extraction.

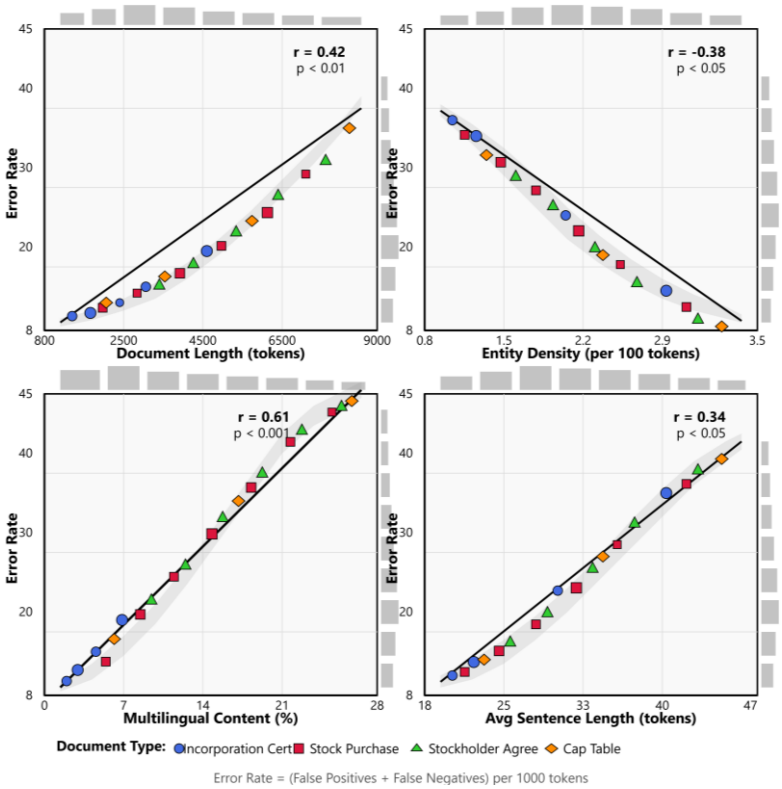

Figure 3. Error Rate Correlation with Document Complexity Features.

The error correlation visualization presents a 2×2 panel grid with four scatter plots examining relationships between document features and extraction error rates. Each panel shows the error rate (false positives + false negatives per 1000 tokens) on the y-axis, plotted against specific document features on the x-axis. The top-left panel examines document length (x-axis: total token count, range 800-9000) versus error rate. Each point represents one test document colored by type: incorporation certificates (blue circles), stock purchase agreements (red squares), stockholder agreements (green triangles), and cap tables (orange diamonds). Point sizes scale 80-200 pixels proportional to entity count.

A fitted regression line in black with 95% confidence interval shading in light gray illustrates positive correlation ($r=0.42$, $p<0.01$, displayed in upper-right corner) indicating longer documents experience elevated error rates potentially due to accumulated context confusion. Top-right panel analyzes entity density (x-axis: entities per 100 tokens, range 0.8-3.5) versus error rate revealing negative correlation ($r=-0.38$, $p<0.05$). Documents with concentrated entity mentions achieve higher extraction accuracy, likely because of repetitive phrasing patterns in entity-dense sections. Scatter points follow same color-coding.

The bottom-left panel plots the percentage of multilingual content (x-axis: percentage of non-ASCII characters, range 0-28%) versus error rate, demonstrating a strong positive correlation ($r=0.61$, $p<0.001$) and confirming that Chinese-English mixed documents pose substantial extraction challenges. The regression line has a steeper slope than the other panels, indicating that multilingual content is the strongest predictor of extraction difficulty. The bottom-right panel examines structural complexity, measured as average sentence length (x-axis: mean tokens per sentence, range 18-47), versus error rate, showing a moderate positive correlation ($r=0.34$, $p<0.05$), indicating that complex sentence

structures impede entity recognition. Marginal histograms along the top and right edges display feature distributions across the test set, with 15 bins per histogram rendered as light-gray bars with black outlines.

5. Discussion and Conclusion

5.1. Key Findings and Implications

5.1.1. Best Performing Methods and Configurations

Experimental findings demonstrate that Legal-BERT is an optimal approach, achieving an 87.3% F1 score through legal-domain specialization. The performance advantage over general-purpose transformers quantifies legal pre-training benefits at 3.6-5.2 percentage points with gains in improved recall for legal terminology and complex structures. The comparative disadvantage of traditional approaches demonstrates the limitations of feature engineering and recurrent architectures. The 8.4-12.7-point F1 gaps exceed those in the general-domain NER, suggesting that legal text particularly benefits from bidirectional transformer attention and large-scale pre-training.

Training efficiency showed that transformers required 4-5 hours, compared to sub-hour traditional methods, representing a 10× increase in computational time. Inference measurements documented CRF baselines processed documents 10× faster, creating practical tradeoffs between accuracy and efficiency. These considerations suggest that hybrid architectures combining fast traditional screening with transformer refinement may optimize the accuracy-efficiency frontier.

5.1.2. Practical Recommendations for M&A Due Diligence

Implementation recommendations emphasize domain-adapted pre-trained models as a foundation for production. Legal-BERT or a comparable specialized transformer should serve as the default, with FinBERT as a fallback. Investment in domain-specific pre-training delivers measurable improvements justifying additional computational resources. Active learning strategies can enhance dataset efficiency by selectively annotating documents with characteristics associated with elevated error rates. Documents featuring multilingual content, complex structures, or unfamiliar types should receive annotation prioritization.

Hybrid pipelines combining automated predictions with targeted manual review of low-confidence extractions balance accuracy with efficiency. Confidence-based filtering, retaining high-probability predictions for automated processing while routing uncertain cases to human review, enables quality-assured outputs without the need for comprehensive manual review. Entity resolution and relationship extraction modules should complement core NER, addressing limitations in handling complex ownership structures and delivering structured outputs that support downstream analytical tasks.

5.2. Limitations and Future Work

5.2.1. Data and Generalization Limitations

Dataset limitations include a modest corpus size (127 documents), restricted geographic coverage (U.S.-centric, with limited international representation), and a technology-sector concentration. Model performance on alternative industries remains invalidated. Generalization to international structures governed by non-U.S. frameworks requires additional validation due to potential differences in documentation. Temporal coverage spanning 2019-2023 may not capture evolving practices or shifts in terminology. Legal document language exhibits gradual drift as drafting conventions adapt to new structures and regulatory requirements.

5.2.2. Directions for Future Research

Relationship extraction integration represents a critical extension that enables the construction of a structured ownership graph. Joint entity-relation models that simultaneously identify shareholders and corresponding equity quantities would better

capture relational semantics. Graph neural architectures operating over document-level entity graphs could model ownership patterns. Cross-lingual model development for Chinese, Korean, and Japanese entities would enhance the applicability of Asian transactions. Multilingual pre-trained models fine-tuned on mixed-language documents could improve code-switched entity handling. Few-shot learning approaches could address rare-entity-type challenges, such as special rights provisions that appear infrequently in the training data.

5.3. Conclusion

5.3.1. Summary of Contributions

This research provides a systematic methodology for comparing ownership structure extraction, addressing critical automation needs in corporate transactions. The study constructs an annotated dataset of authentic governance documents providing benchmark resources for evaluating extraction approaches in corporate law contexts. A comprehensive comparison of six NER methodologies quantifies the performance impact of architectural sophistication and domain specialization. Experimental findings establish that Legal-BERT achieves optimal performance at 87.3% F1, representing substantial improvements over traditional approaches and moderate advantages over general-purpose transformers.

Entity-type-specific analysis reveals differential performance across ownership categories, with shareholder identification achieving the highest accuracy, and special-rights extraction presenting persistent challenges. Error analysis identifies multilingual recognition, complex structures, and low-frequency types as primary obstacles. The work provides actionable guidance for deploying automated extraction, emphasizing domain-specialized models, active learning strategies, and confidence-based workflows. Relationship extraction, integration, and multilingual development emerge as priority areas for advancing automated due diligence capabilities that support efficient cross-border M&A activity.

References

1. A. Shah, A. Gullapalli, R. Vithani, M. Galarnyk, and S. Chava, "FiNER-ORD: Financial named entity recognition open research dataset," *arXiv preprint arXiv:2302.11157*, 2023.
2. X. Zhang, X. Luo, and J. Wu, "A Roberta-globalpointer-based method for named entity recognition of legal documents," In *2023 International Joint Conference on Neural Networks (IJCNN)*, 2023, pp. 1-8. doi: 10.1109/ijcnn54540.2023.10191275
3. L. Hillebrand, T. Deußer, T. Dilmaghani, B. Kliem, R. Loitz, C. Bauckhage, and R. Sifa, "KPI-BERT: A joint named entity recognition and relation extraction model for financial reports," In *2022 26th International Conference on Pattern Recognition (ICPR)*, 2022, pp. 606-612.
4. B. Aejas, A. Belhi, H. Zhang, and A. Bouras, "Deep learning-based automatic analysis of legal contracts: A named entity recognition benchmark," *Neural Computing and Applications*, vol. 36, no. 23, pp. 14465-14481, 2024. doi: 10.1007/s00521-024-09869-7
5. I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The muppets straight out of law school," *arXiv preprint arXiv:2010.02559*, 2020. doi: 10.18653/v1/2020.findings-emnlp.261
6. K. Guo, T. Jiang, and H. Zhang, "Knowledge graph enhanced event extraction in financial documents," In *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 1322-1329. doi: 10.1109/bigdata50022.2020.9378471
7. F. Ariai, J. Mackenzie, and G. Demartini, "Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges," *ACM Computing Surveys*, 2024. doi: 10.1145/3777009
8. D. Hendrycks, C. Burns, A. Chen, and S. Ball, "CUAD: An expert-annotated NLP dataset for legal contract review," *arXiv preprint arXiv:2103.06268*, 2021.
9. S. Skylaki, A. Oskooei, O. Bari, N. Herger, and Z. Kriegman, "Named entity recognition in the legal domain using a pointer generator network," *arXiv preprint arXiv:2012.09936*, 2020.
10. R. Lu, and L. Li, "Named entity recognition method of Chinese legal documents based on parallel instance query network," *International Journal of Digital Crime and Forensics (IJDCF)*, vol. 16, no. 1, pp. 1-19, 2024. doi: 10.4018/ijdcf.367470
11. H. Hamad, A. K. Thakur, N. Kolleri, S. Pulikodan, and K. Chugg, "FIRE: A dataset for financial relation extraction," In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 3628-3642. doi: 10.18653/v1/2024.findings-naacl.230
12. E. Leitner, G. Rehm, and J. M. Schneider, "A dataset of German legal documents for named entity recognition," In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 4478-4485.

13. Y. Chen, Y. Sun, Z. Yang, and H. Lin, "Joint entity and relation extraction for legal documents with legal feature enhancement," In *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 1561-1571. doi: 10.18653/v1/2020.coling-main.137
14. Y. Zhang, and H. Zhang, "FinBERT-MRC: Financial named entity recognition using BERT under the machine reading comprehension paradigm," *Neural Processing Letters*, vol. 55, no. 6, pp. 7393-7413, 2023. doi: 10.1007/s11063-023-11266-5
15. S. F. Mohsin, S. I. Jami, S. Wasi, and M. S. Siddiqui, "An automated information extraction system from the knowledge graph based annual financial reports," *PeerJ Computer Science*, vol. 10, p. e2004, 2024. doi: 10.7717/peerj-cs.2004