Development of an Enterprise-Grade Contract Understanding System

A. Agarwal[2], L. Chiticariu[3], P. Chozhiyath Raman[3], M. Danilevsky[1], D. Ghazi[5]*, A. Gupta[2], S. Guttula[2], Y. Katsis[1], R. Krishnamurthy[3], Y. Li[1], S. Mudgal[3], V. Munigala[2], N. Phan[6]*, D. Sonawane[3], S. Srinivasan[3], S. Thitte[3], M. Vasa[3], R. Venkatachalam[3], V. Yaski[4]*, H. Zhu[1]

[1] IBM Research - Almaden [2] IBM Research - India [3] IBM Data and AI
[4] Amazon [5] Google LLC [6] Visa

{arvagarw, ankushgupta, shguttul, vmunig10}@in.ibm.com; {chiti, pchozhi, mdanile, rajase, yunyaoli, srthitte, mitesh.vasa, venkatra, huaiyu}@us.ibm.com; {yannis.katsis, shubham.mudgal, dhaval.sonawane1, sneha.srinivasan1130}@ibm.com; dimanghazi@google.com; thefryingphan@gmail.com; vinitha.yaski@gmail.com

Abstract

Contracts are arguably the most important type of business documents. Despite their significance in business, legal contract review largely remains an arduous, expensive and manual process. In this paper, we describe the Transparent and Expert Contract Understanding System (TECUS): a commercial system designed and deployed for contract understanding and used by a wide range of enterprise users for the past few years. We reflect on the challenges and design decisions when building TECUS. We also summarize the data science life cycle of TECUS and share lessons learned.

1 Introduction

A *contract* is an agreement between businesses and/or individuals to create mutual obligations enforceable by law (Cornell Law School). Written contracts are also used by companies to safeguard their resources and as such, legal advice is sought prior to participating in a binding contract. Currently, legal review remains an arduous and expensive process. For instance, a procurement contract requires 5 hours of legal review on average, contributing to thousands of dollars in total cost (Cummins, 2017).

While contract reviewing is a well-established legal process, building an enterprise-grade system for *Contract Understanding* (CU) to facilitate this process poses three major challenges:

C1: Model CU as an NLP Problem. CU does not have a corresponding standard NLP definition. The underlying processes and the associated requirements for CU need to be well understood to translate it to concrete NLP tasks.

Table 1: Example Contract Understanding Use-cases

Context	Application
Quote to Cash	Identify non-standard, risky terms
Accounts Receivable	Prevent leakage, improve cash-flow
Procurement	Analyze numerous contracts in negotiation
Global Accounting	Assist with numerous compliance checklists
Mergers & Acquisitions	Identify early termination notice period, penalty amount etc.

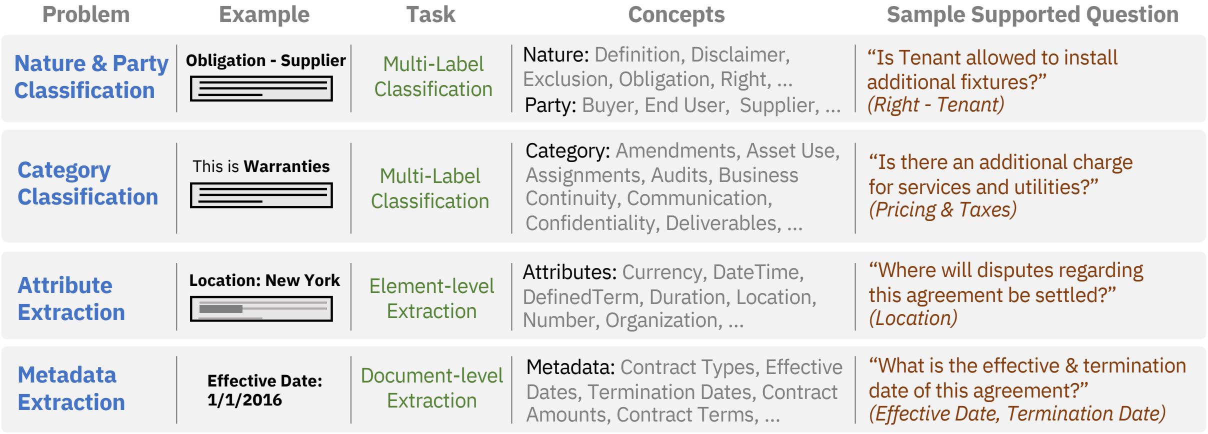
C2: Lack of Representative Data. Contracts, while often proprietary, also vary significantly across domains and businesses. Thus, one cannot assume the presence of representative contracts towards building models. Moreover, Subject Matter Experts (SMEs) qualified to label ground truth are expensive[1]. As such, NLP models may need to be developed with limited non-representative labeled data but still be able to generalize well over previously unseen data.

C3: Need for Model Stability. CU models are integrated into existing business processes to drive decisions. As the models evolve over time (e.g. due to availability of new data, updates to existing labeled data, etc.), users expect the models to behave in a stable manner and produce no surprising results (Kearns and Ron, 1997).

[*] Work done while author was working at IBM.

[1] According to https://www.zippia.com/contract-attorney-jobs/salary/, the average annual salary for a contract attorney is $86,000 ($41.35/hour).

Proceedings of NAACL HLT 2021: IndustryTrack Papers, pages 222–229
June 6–11, 2021. ©2021 Association for Computational Linguistics



Problem	Example	Task	Concepts	Sample Supported Question
Nature & Party Classification	Obligation - Supplier	Multi-Label Classification	Nature: Definition, Disclaimer, Exclusion, Obligation, Right, ... Party: Buyer, End User, Supplier, ...	"Is Tenant allowed to install additional fixtures?" *(Right - Tenant)*
Category Classification	This is Warranties	Multi-Label Classification	Category: Amendments, Asset Use, Assignments, Audits, Business Continuity, Communication, Confidentiality, Deliverables, ...	"Is there an additional charge for services and utilities?" *(Pricing & Taxes)*
Attribute Extraction	Location: New York	Element-level Extraction	Attributes: Currency, DateTime, DefinedTerm, Duration, Location, Number, Organization, ...	"Where will disputes regarding this agreement be settled?" *(Location)*
Metadata Extraction	Effective Date: 1/1/2016	Document-level Extraction	Metadata: Contract Types, Effective Dates, Termination Dates, Contract Amounts, Contract Terms, ...	"What is the effective & termination date of this agreement?" *(Effective Date, Termination Date)*

Figure 1: TECUS' Sub-Problems (see (IBM, b) for complete list of supported concepts)

To overcome the above challenges, we designed and developed the **Transparent and Expert Contract Understanding System (TECUS)**, a commercial system that enables legal professionals to review contracts with minimal effort.

TECUS first models CU as a series of text classification and extraction tasks, defined collaboratively with SMEs, to capture the information that legal experts seek when reviewing contracts.

Second, it leverages SystemT, a state-of-the-art declarative text understanding engine for the enterprise (Chiticariu et al., 2010, 2018), towards developing transparent models on top of syntactic and semantic linguistic features, to mitigate a possible lack of representative labeled data and to satisfy model stability requirements. This approach enables (1) the development of stable models that explicitly capture domain knowledge without requiring large amounts of labeled data or representative samples; and (2) a data science workflow that supports systematic error analysis and incorporation of user feedback towards continuous model improvement (Section 3).

TECUS is available as part of multiple commercial products including IBM Watson® Discovery (IBM, a) and IBM Watson® Compare and Comply [2]. As part of these products, it has been in use by enterprise customers since 2017 to support a variety of contract understanding use-cases (Table 1). While several other commercial offerings, such as Cognitiv+ (Cog), Kira (Kir), LawGeex (Law), LegalSifter (Leg), and Lexion (Lex) use NLP to analyze contracts, their internals are not publicly disclosed. Thus, TECUS is to the best of our knowledge the first commercial automated contract understanding system ever presented to the scientific community in such detail.

In addition to assisting in the understanding of a single contract, as described in this paper, TECUS also allows legal professionals to compare two contracts, identifying similarities and differences along multiple dimensions; another critical task in the contract reviewing process. TECUS models this problem as a clause-level comparison problem, identifying (i) clauses that are identical between two contracts, (ii) clauses that are on the same topic but have changed, and (iii) clauses that appear in one contract but not in the other. The comparison component, similar to the contract understanding component, leverages syntactic and semantic linguistic features provided by SystemT and an associated data science workflow tuned towards a systematic and stable model development. However, for space reasons, this work focuses on single contract analysis, enabled by TECUS' Contract Understanding (CU) component.

2 Modeling CU as an NLP Problem

Working with legal experts, we first define the CU problem as a combination of Multi-class Multi-label[3] Classification and Entity Extraction tasks, as depicted in Figure 1.

Clause Classification. A business contract consists of thousands of sentences, each defining one or more clauses, such as *Obligation*, *Exclusion*, etc. At the core of the legal review process are identi-

[2] IBM and IBM Watson are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide.

[3] The classification problems correspond to multi-label classification, as elements are often complex and cover multiple Categories/Natures/Parties.

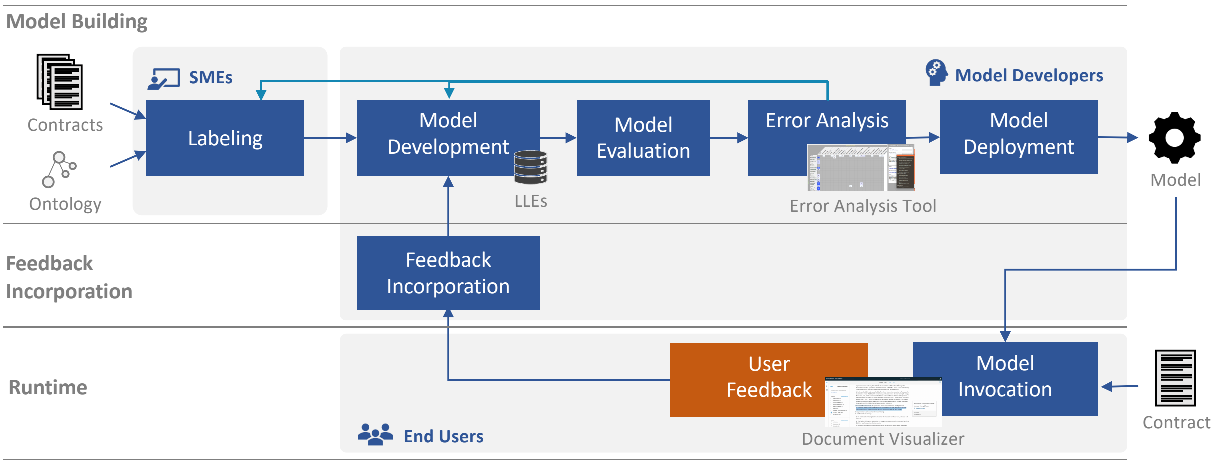


Figure 2: TECUS' Architecture

fying, classifying and reviewing individual clauses to spot potential risks. For example, the sentence

> *"Purchaser will purchase the Assets by a cash payment of FOUR HUNDRED FIFTEEN THOUSAND US DOLLARS."*

is a clause related to *Pricing & Taxes* that describes an *Obligation* for the *Purchaser*. To help legal professionals focus on relevant parts of the contract, we classify contract sentences (henceforth known as *elements* [4]) according to three dimensions of interest to domain experts:

- **Category**: the topic associated with the element, such as *Pricing & Taxes* in our example.
- **Nature**: the action described by the element, such as *Obligation* in our example.
- **Party**: the individual or entity affected by the action, such as *Supplier* in our example.

A consistent ontology (shown in Figure 1) was defined in the early stages of the project, in collaboration with SMEs via an iterative process, and reflecting the prevailing views of legal experts. However, to also accommodate users who adopt slight variations of the definitions (which we discovered can be common due to the subtle nature of legal terms), users can also customize TECUS through user feedback, as described in Section 3.3.

Attribute and Metadata Extraction. In addition to classifying elements, legal teams are also interested in extracting entities of particular importance to corporate law. These fall into two categories:

- **Attributes**: general entities of interest, such as *Organizations* and *Persons* involved in a contract, *Dates*, *Locations* and *Currencies*. Attributes are extracted from individual elements.
- **Contract Metadata**: document-level legal entities of interest, such as the *Effective Dates*, *Termination Dates*, and *Contract Amounts*. Contract Metadata are extracted from across a contract, and are thus applicable to the entire contract.

3 System Overview

As can be seen in Figure 2, TECUS consists of three main components, which we subsequently describe in detail: *Runtime*, *Model Building*, and *Feedback Incorporation*.

3.1 Runtime

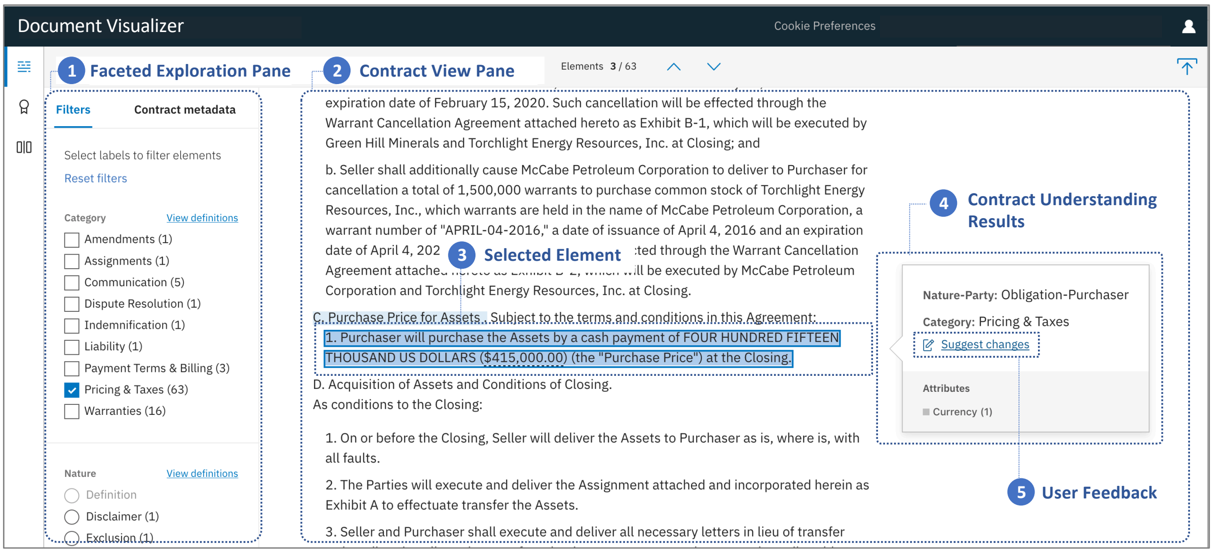
As shown in Figure 3, users can analyze their contracts by interacting with TECUS's *Document Visualizer*, via the following two panes:

The *Faceted Exploration Pane* (#1) allows users to quickly acquire an overview of the contract's contents and drill down into specific categories/natures/parties of interest. In our example, a user interested in *Pricing & Taxes*, focuses on such clauses by selecting the corresponding checkbox.

The *Contract View Pane* (#2) allows users to see the selected elements within the contract (#3) and for each of them inspect the *Category/Nature/Party* and *Attributes* identified by CU (#4). It also includes a *Metadata View* showing the metadata extracted from the contract, such as *Contract Amounts*, *Effective Dates*, *Termination Dates*, etc. (omitted in the interest of space).

At any point in time, users can provide feedback on the CU results through the "Suggest changes" feature (#5). User feedback is then further analyzed and incorporated as described in Section 3.3.

[4] TECUS supports classification of elements beyond sentences, including bulleted list items and table content, which often appear in contracts.



Document Visualizer

Cookie Preferences

Elements 3 / 63

Filters Contract metadata

Select labels to filter elements

Reset filters

Category View definitions

- □ Amendments (1)
- □ Assignments (1)
- □ Communication (5)
- □ Dispute Resolution (1)
- □ Indemnification (1)
- □ Liability (1)
- □ Payment Terms & Billing (3)
- ☑ Pricing & Taxes (63)
- □ Warranties (16)

Nature View definitions

- ○ Definition
- ○ Disclaimer (1)
- ○ Exclusion (1)

expiration date of February 15, 2020. Such cancellation will be effected through the Warrant Cancellation Agreement attached hereto as Exhibit B-1, which will be executed by Green Hill Minerals and Torchlight Energy Resources, Inc. at Closing; and

b. Seller shall additionally cause McCabe Petroleum Corporation to deliver to Purchaser for cancellation a total of 1,500,000 warrants to purchase common stock of Torchlight Energy Resources, Inc., which warrants are held in the name of McCabe Petroleum Corporation, a warrant number of "APRIL-04-2016," a date of issuance of April 4, 2016 and an expiration date of April 4, 202 :ted through the Warrant Cancellation Agreement attache .ill be executed by McCabe Petroleum Corporation and Torchlight Energy Resources, Inc. at Closing.

C. Purchase Price for Assets . Subject to the terms and conditions in this Agreement:

1. Purchaser will purchase the Assets by a cash payment of FOUR HUNDRED FIFTEEN THOUSAND US DOLLARS ($415,000.00) (the "Purchase Price") at the Closing.

D. Acquisition of Assets and Conditions of Closing.
As conditions to the Closing:

1. On or before the Closing, Seller will deliver the Assets to Purchaser as is, where is, with all faults.

2. The Parties will execute and deliver the Assignment attached and incorporated herein as Exhibit A to effectuate transfer the Assets.

3. Seller and Purchaser shall execute and deliver all necessary letters in lieu of transfer

Nature-Party: Obligation-Purchaser

Category: Pricing & Taxes

Suggest changes

Attributes

Currency (1)

Figure 3: Reviewing a Contract through TECUS' Document Visualizer

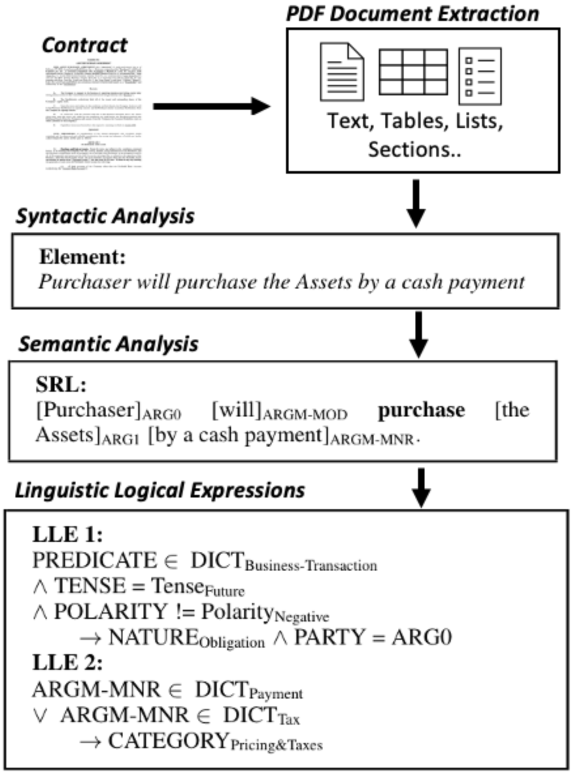

Contract
PDF Document Extraction
Text, Tables, Lists, Sections..
Syntactic Analysis
Element:
Purchaser will purchase the Assets by a cash payment
Semantic Analysis
SRL:
[Purchaser]ARG0 [will]ARGM-MOD purchase [the Assets]ARG1 [by a cash payment]ARGM-MNR.
Linguistic Logical Expressions
LLE 1:
PREDICATE ∈ DICT Business-Transaction
∧ TENSE = Tense Future
∧ POLARITY != Polarity Negative
→ NATURE Obligation ∧ PARTY = ARG0
LLE 2:
ARGM-MNR ∈ DICT Payment
∨ ARGM-MNR ∈ DICT Tax
→ CATEGORY Pricing&Taxes

Figure 4: Contract Understanding Components

3.2 Model Building

3.2.1 Model Development

Figure 4 illustrates the sequence of model components used by TECUS to accomplish the tasks in Figure 1. TECUS uses declarative models towards both classification and extraction tasks [5]:

(1) Once text and document structures such as lists, sections, tables etc. are extracted from a contract PDF document [6], sentences/elements (for classification) and tokens/phrases (for extraction) are identified using syntactic analysis.

(2) Next, extended Semantic Role Labels (SRL) (Palmer et al., 2010), provided by SystemT (Chiticariu et al., 2010, 2018) [7]are identified in elements. As shown in Figure 4, SRL captures *who did what to whom, when, where, and how* from the example element.

(3) A collection of logical formulae, called *Linguistic Logical Expressions (LLEs)*, are constructed using these SRLs to perform logical reasoning using linguistic patterns in contracts, similar to reasoning by legal experts. For instance, the two LLEs [8] in Figure 4 identify the Category, Nature, and Party concerning the example element.

Each classification model in TECUS consists of a collection of such LLEs. Such a model not only yields a transparent understanding of a contract along the concepts outlined in Figure 1, it is also uniquely positioned to handle the challenges outlined in Section 1 for the following reasons:

Generalizability. LLEs are manually built by model developers on top of SRL[9], to explicitly

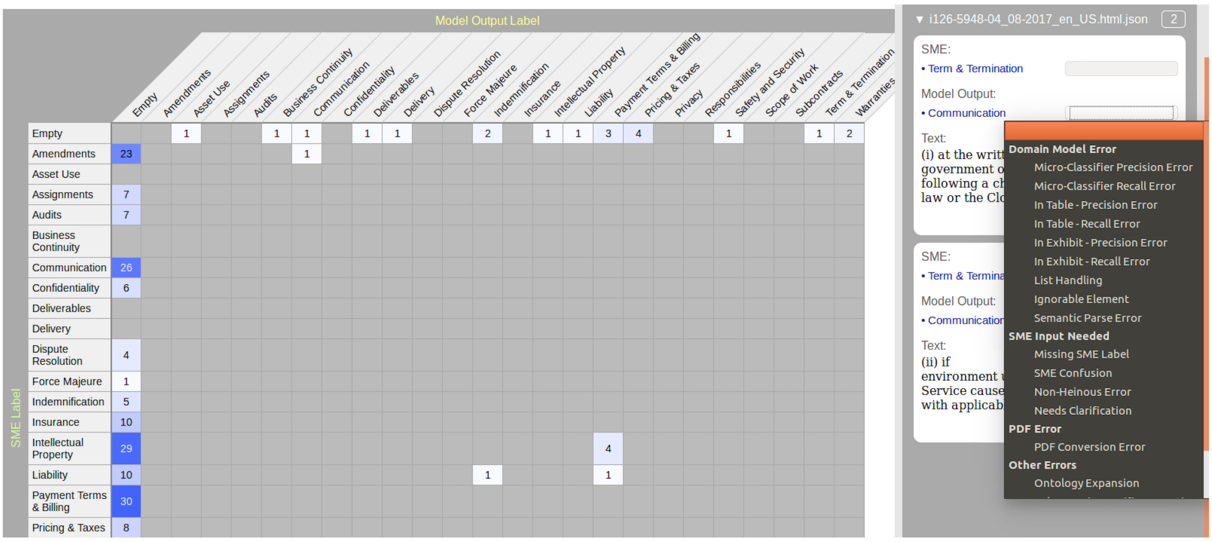
[5] Here, we focus on classification due to its challenging aspects. Attribute and Metadata extraction are performed using entity extraction, enabled by SystemT.

[6] Here, we use PDF as the business document format for ease of exposition. The presented techniques apply also to other document formats, such as Microsoft Word.

[7] SystemT also provides additional information, such as tense and voice; please refer to (Zhu et al., 2019) for details.

[8] Simplified from the actual product LLEs for readability.

[9] Potentially aided by machine learning (Sen et al., 2019)



	Empty	Amendments	Asset Use	Assignments	Audits	Business Continuity	Communication	Confidentiality	Deliverables	Delivery	Dispute Resolution	Force Majeure	Indemnification	Insurance	Intellectual Property	Liability	Payment Terms & Billing	Pricing & Taxes	Privacy	Responsibilities	Safety and Security	Scope of Work	Subcontracts	Term & Termination	Warranties
Empty			1			1	1		1	1			2		1	1	3	4			1			1	2
Amendments	23						1																		
Asset Use																									
Assignments	7																								
Audits	7																								
Business Continuity																									
Communication	26																								
Confidentiality	6																								
Deliverables																									
Delivery																									
Dispute Resolution	4																								
Force Majeure	1																								
Indemnification	5																								
Insurance	10																								
Intellectual Property	29																4								
Liability	10												1				1								
Payment Terms & Billing	30																								
Pricing & Taxes	8																								

▼ i126-5948-04_08-2017_en_US.html.json 2

SME:
• Term & Termination

Model Output:
• Communication

Text:
(i) at the writ
government o
following a ch
law or the Cl

SME:
• Term & Termina

Model Output:
• Communication

Text:
(ii) if
environment
Service cause
with applicab

Figure 5: Analyzing CU Errors through the ModelLens Error Analysis Tool

capture domain knowledge. Each LLE reflects patterns from not only the documents used during the development process, but also unseen contracts where similar semantic patterns appear. As a result, the CU model generalizes much better to yet unseen contracts than state-of-the-art black-box models (see Section 4 for more details).

Enabling systematic model improvement workflow. Use of LLEs enable a fine-grained association of CU model output with highly specific, lower-level constructs of the model. This transparency allows a team of developers to make localized updates and develop models with stable and explainable behavior, aided by a carefully created data science workflow of *model evaluation*, *error analysis* and *feedback incorporation*, as described next.

3.2.2 Model Evaluation

As the CU model in TECUS evolves over time, it is regularly evaluated for: (1) quality, in terms of precision, recall and accuracy towards both Nature-Party and Category classification tasks, and (2) performance, in terms of throughput, memory consumption and behavior profile.

We measure model quality over in-domain and out-of-domain data split into the usual train (dev) and test (hold-out) subsets. Similarly, we profile runtime performance upon multiple in-domain and out-of-domain sets, allowing developers to preemptively rectify potentially problematic runtime behaviors, prior to deployment.

Beyond the typical global measurements, the transparent nature of the CU model permits evaluation at finer granularity: across classes, per-class and per-LLE towards both quality and runtime performance. Such detailed model evaluation along with a systematic model improvement workflow together enable TECUS to provide reliable guarantees of consistency and robustness of its results.

3.2.3 Error Analysis

While evaluation provides an overview of model performance, model improvement requires delving deeper and analyzing individual errors to understand their root causes and inform further model development efforts (Ribeiro et al., 2020).

TECUS supports root cause identification of errors through the *ModelLens* error analysis tool shown in Figure 5 and the associated error analysis workflow (Katsis and Wolf, 2019). Specifically, ModelLens allows model developers to perform the following error analysis tasks:

(1) Acquire a high-level overview of the errors through a confusion matrix that depicts the types of misclassifications made by the model, to help prioritize errors for further analysis.

(2) Inspect erroneous instances in context. For a chosen misclassification type, developers can drill down and inspect all elements that exhibit it (shown on the right side of the screen). For each element, ModelLens also provides additional context, such as the surrounding text, the SRL output, and the provenance of the model output, to help model developers identify the error root cause.

(3) Annotate errors with their root causes. Once a developer identifies the root cause of an error, they can record it through the drop-down next to the

corresponding label.

This error analysis process classifies errors based on their root causes, separating true model errors from other errors that have to be treated differently (e.g., labeling errors, errors from preceding models, such as PDF conversion errors, and others). Additionally, ModelLens exploits the transparent nature of the model to allow developers to identify specific LLEs that need to be revised to address model errors. Moreover, by providing contextual information for each error, it also assists in identifying additional linguistic patterns that could be translated into new LLEs.

3.3 Feedback Incorporation

User feedback is essential for TECUS: First, it enables model improvement, by communicating to the development team cases not captured by the current model. This is especially important given the lack of representative labeled data discussed in Section 1. Second, it allows the customization of models. Custom models allow TECUS to adapt to the needs of individual customers, who may adopt slightly different definitions of the Category/Nature/Party classes from our SMEs, as described in Section 2. Feedback is enabled by the following human-in-the-loop process:

(1) Users review model results and suggest the exclusion of incorrect labels or the inclusion of missing labels through the Document Visualizer.

(2) The system locates other elements that share a similar linguistic pattern (i.e., LLE) with the ones on which feedback was provided, and asks the user whether they would like to propagate the suggested label updates to those. This capability is to reduce user efforts in providing feedback.

(3) The system associates user feedback to the corresponding LLEs, allowing model changes to be localized to a small part of the model.

The localized nature of the changes enable the model to remain stable over time; in contrast, black box models may regress in unexpected ways when globally retrained over time with additional ground truth data. Results on model stability and the effectiveness of feedback incorporation are presented in the next section.

4 Results & Discussion

We next present evaluation results, showing how TECUS addresses the challenges discussed above.

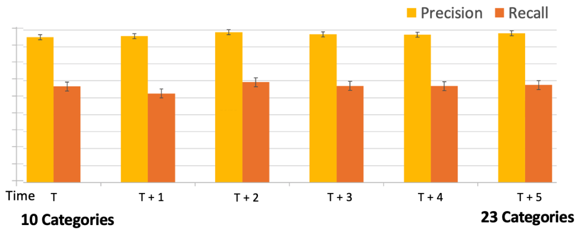


Figure 6: Model Stability with Increasing Complexity

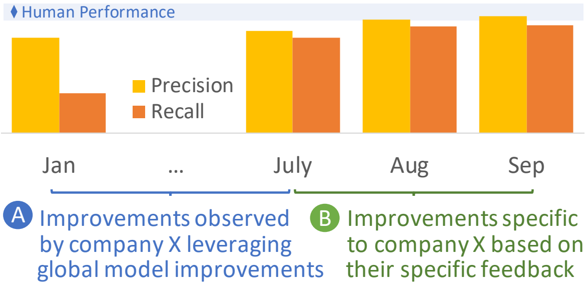


Figure 7: Effectiveness of Feedback Incorporation

Model generalizability. To verify our intuition that the transparent nature of the CU model helps it generalize to unseen contracts, we compare it to state of the art black box ML models. In our experiments, when trained on procurement contracts (PCs) sourced from within IBM (in-domain) and tested on PCs sourced randomly from the web (out-of-domain), the CU model significantly outperforms the alternatives, with 1.3x, 2.09x, and 2.58x higher micro-$F1$ score over a Bidirectional Long Short-Term Memory (BiLSTM), Convolutional Neural Network (CNN), and Logistic Regression (LR) model, respectively.

For reference, prior state-of-the-art results of nature identification in contracts (Chalkidis et al., 2018) and financial legislation (Neill et al., 2017) were based on different BiLSTM-based architectures, which we have found in our experiments to perform well on contracts similar to the ones on which they were trained (e.g., contracts that follow similar templates) but generalize poorly to other unseen contracts.

Model stability. To verify the stability of model performance over time, we capture in Figure 6 the CU model's precision and recall on category classification across six consecutive development sprints (each two weeks long). During this time period, the development team added support for additional categories, increasing the supported categories from 10 to 23. Despite a quick addition of new categories (with 2.25 new categories added per sprint on average), the model's quality remained stable across all categories, old and new. This can be attributed (a) to the transparent nature of the model, which allows changes to be localized and (b) to the data science process which allows quick additions of new categories without compromising on quality.

Effectiveness of feedback incorporation. To verify the effectiveness of TECUS's feedback incorporation mechanism, we capture in Figure 7 the CU model's precision and recall for category classification on data of interest for an individual customer X over 9 months. During this period the model development team incorporated two types of feedback: in the first few months (January-July), they leveraged feedback solely from other customers, while in the last few months (July-Sep), they incorporated focused feedback from customer X. As shown in the chart, customer X benefited both from the feedback given by other customers, as well as its own feedback, which further improved quality. Moreover, the model quality increased consistently towards human performance (calculated internally based on evaluation of inter-annotator agreement among SMEs).

5 Conclusion

We have presented TECUS, a commercial system that effectively assists and supplements legal experts in understanding and reviewing contracts. TECUS' effectiveness is based on (a) the transparent nature of the CU model, comprised of Linguistic Logical Expressions on top of SRL, which in turn enables (b) a systematic data science workflow towards swift yet stable model development. This leads to models that can be developed with limited, non-representative labeled data and remain stable and predictable over time; traits that are essential not just in the contract understanding domain but the wider legal domain as well. Finally, while the system was developed for the CU problem, we believe that its design and associated insights could inform efforts in other areas that pose similar requirements of generalizability and stability.

References

Cognitiv+ (Accessed: 2021-04-09). http://www.cognitivplus.com.

a. IBM Watson Discovery (Accessed: 2021-04-09). https://www.ibm.com/cloud/watson-discovery.

b. IBM Watson Discovery: Understanding Contract Analysis (Accessed: 2021-04-09). https://cloud.ibm.com/docs/discovery-data?topic=discovery-data-contract_parsing.

Kira (Accessed: 2021-04-09). https://kirasystems.com.

LawGeex (Accessed: 2021-04-09). https://www.lawgeex.com.

LegalSifter (Accessed: 2021-04-09). https://www.legalsifter.com.

Lexion (Accessed: 2021-04-09). https://lexion.ai.

Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2018. Obligation and Prohibition Extraction Using Hierarchical RNNs. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 254–259, Melbourne, Australia. Association for Computational Linguistics.

Laura Chiticariu, Marina Danilevsky, Yunyao Li, Frederick Reiss, and Huaiyu Zhu. 2018. SystemT: Declarative Text Understanding for Enterprise. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 76–83. Association for Computational Linguistics.

Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Sriram Raghavan, Frederick R. Reiss, and Shivakumar Vaithyanathan. 2010. SystemT: An Algebraic Approach to Declarative Information Extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. ACL '10*, pages 128–137. Association for Computational Linguistics.

Cornell Law School (Accessed: 2021-04-09). Contract. https://www.law.cornell.edu/wex/contract.

Tim Cummins. 2017. Cost of processing a basic contract soars to $6900 (accessed: 2021-04-09). https://blog.lawgeex.com/contractcosts/.

Yannis Katsis and Christine T. Wolf. 2019. ModelLens: an interactive system to support the model improvement practices of data science teams. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, CSCW '19, page 9–13, New York, NY, USA. Association for Computing Machinery.

Michael Kearns and Dana Ron. 1997. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, COLT '97, page 152–162, New York, NY, USA. Association for Computing Machinery.

James O' Neill, Paul Buitelaar, Cecile Robin, and Leona O' Brien. 2017. Classifying Sentential Modality in Legal Language: A Use Case in Financial Regulations, Acts and Directives. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, ICAIL '17, page 159–168, New York, NY, USA. Association for Computing Machinery.

Martha Palmer, Daniel Gildea, and Nianwen Xue. 2010. *Semantic Role Labeling*. Synthesis Lectures on Human Language Technology Series. Morgan and Claypool.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Prithviraj Sen, Yunyao Li, Eser Kandogan, Yiwei Yang, and Walter Lasecki. 2019. HEIDL: Learning linguistic expressions with deep learning and human-in-the-loop. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 135–140, Florence, Italy. Association for Computational Linguistics.

Huaiyu Zhu, Yunyao Li, and Laura Chiticariu. 2019. Towards universal semantic representation. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 177–181, Florence, Italy. Association for Computational Linguistics.