AIAS-2019

Proceedings of the First Workshop on AI and the Administrative State

University of Montreal, Quebec, Canada
17 June 2019
Held in conjunction with ICAIL 2019
https://icail2019-cyberjustice.com/

L. Karl Branting (Editor)

Preface

Much of the work of modern governance, including delivery of public services, adjudication of entitlement to public benefits, and enforcement of legal mandates, is performed by administrative agencies implementing statutes, regulations, and other authoritative legal sources expressed in complex, interconnected texts. Understanding and complying with these rules is challenging for agencies, citizens, rule-drafters, and attorneys alike.

Recent advances in AI, Machine Learning, Human Language Technology, Network Science, and Human Factors analysis offer promising new approaches to improving the ability of all stakeholders, including agencies themselves, to operate within this complex regulatory environment. The scale of administrative states means that the benefits of automation have very high potential impact, both in improvements to government processes and in the delivery of services and benefits to citizens. At the same time, the black-box nature of many automated decision-making systems, particularly sub-symbolic AI components such as those generated by machine learning algorithms, can create considerable tension with the norms of transparency, accountability, and reason-giving that typically govern administrative action. Explainable, responsible, and trustworthy AI is vital for addressing these factors.

TABLE OF CONTENTS

Segmentation of Rulemaking Documents for Public Notice-and-Comment Process Analysis

Anna Belova*
abelova@alumni.cmu.edu
Carnegie Mellon University
Pittsburgh, PA

Matthias Grabmair
mgrabmai@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA

Eric Nyberg
ehn@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA

ABSTRACT

We evaluate feasibility of automated identification of comment discussion passages and comment-driven proposed rule revisions in the US Environmental Protection Agency's (EPA's) rulemaking documents. We have annotated a dataset of final rule documents to identify all spans in which EPA discusses and evaluates the merits of public comments received on its proposed rules, and present lessons learned from the annotation process. We implement several baseline supervised discourse segmentation models that combine classic linear learners with sentence representations using handcrafted features as well as Bidirectional Encoder Representations from Transformers (BERT). We observe good agreement on annotation comment discussions and our models achieve a classification F1 of 0.73. Public comment dismissals and rule revisions are substantially harder to annotate and predict, leading to lower agreement and model performance. Our work contributes a dataset and a baseline for a novel discourse segmentation task of identifying public comment discussion and evaluation by the receiving agency.

CCS CONCEPTS

• **Applied computing** → **Law**; **E-government**; *Annotation*; • **Computing methodologies** → *Discourse, dialogue and pragmatics*; Neural networks; Learning in probabilistic graphical models.

KEYWORDS

datasets, supervised discourse segmentation, public notice-and-comment process, SVM, CRF, BERT

1 INTRODUCTION

Government agencies are created by the legislatures worldwide to regulate social, economic, and political aspects of people's lives. These agencies belong to the executive branch of the government, yet they create legally enforceable regulations and rules that implement broad legislation. In the US, public notice-and-comment processes have become an important venue for influencing social and economic policy. In that, US agencies publish proposed rules in the Federal Register (FR) and all interested parties are given an opportunity to comment. Agency regulatory proposals receive public feedback from individuals, businesses, organized groups (of individuals or businesses), and other agencies. Comments represent heterogeneous interests in particular regulatory outcomes. The agency is not obliged to react to each individual received comment. However, it has to respond to comments that raise significant issues with the proposed rule and, if the points raised have merit, may substantively revise of the rulemaking document. The final rule document is published in the FR and contains the discussion of submitted comments, or points to other documents in the docket that address concerns raised in the comments.

The online forum for the US public notice-and-comment process—`regulations.gov`—was launched in January 2003, as part of the US eRulemaking program established as a cross-agency E-Gov initiative under Section 206 of the 2002 E-Government Act (H.R. 2458/S. 803). In this collection, all documents pertaining to the development of a particular rule are compiled in a regulatory docket. A typical docket contains a proposed rule document, many public comment documents, and a final/revised rule document.[1] As such, `regulations.gov` provides a testbed for study of the public notice-and-comment discourse in the US.

In this work, we focus on (1) identifying spans in the final rule documents that contain the agency's discussion of the public comments it received, and (2) classifying those spans as being either dismissals of the commenter claims or revisions of the proposed regulations prompted by the comment. In that, we analyze 353 US Environmental Protection Agency (EPA) regulations proposed in January 2003 or later, and finalized as of March 2018.[2]

Our work contributes a dataset[3] and a baseline for a novel discourse segmentation task of identifying public comment discussion and evaluation by the receiving agency. Automatic detection of comment discussion passages in the rulemaking documents could improve the efficiency of regulatory review conducted by experts at a number of organizations, including the US Office of Information and Regulatory Affairs, regulatory agencies, and other stakeholders of the regulatory process. In addition, segmentation of regulatory discourse is the first step bringing agency's narrative deliberations in the study of bureaucratic politics and decision making (e.g., regulatory capture theory) by economists and political scientists [37], which to date has relied on structured data generated by surveys and administrative record-keeping (e.g. permitting, inspections).

*Corresponding author

In: Proceedings of the First Workshop on AI in the Administrative State, June 17, 2019, Montréal, QC, CA.*AIAS'19, June 17, Montréal, CA*

[1] Other documents, such as transcripts of public hearings, technical support documents, detailed comment response documents, copies of pertinent scientific papers, e-mails and other correspondence, may also be included. Finally, a docket may also contain tabular data and software source code used to produce analytical results.

[2] We have chosen to focus on EPA because this agency published the most rules (∼ 20% of all rule documents) and received the most comment submissions (∼ 10% of all comment documents) in `regulations.gov` during the studied time period.

[3] The data and code are available at https://github.com/mug31416/PubAdminDiscourse.git

2 RELATED WORK

In the peer-reviewed literature, discussion of e-rulemaking benefits, challenges, and related artificial intelligence (AI) methods began in the early 2000s [9]. Over a decade later, surveys by [7] and [37] describe several e-rulemaking initiatives that involved successful applications of AI. One line of e-rulemaking research has focused on tasks relevant to management of massive amount of public comments received by agencies (e.g., [58], [31], [52]). Another line of research, conducted as part of Cornell University's RegulationRoom project, has focused on tools to improve the quality of public discourse around rulemaking (e.g., [41], [46]). Research on the text of rules developed by agencies has mostly focused on the search for similar rules in the FR [33], rather than segmentation of the comment-related discourse in the rule documents.

Prior to launch of `regulations.gov`, work on e-rulemaking used several rule-specific comment collections that were either shared by the agencies—EPA, Fish and Wildlife Service (FWS)—or gathered as part of the RegulationRoom experiments in collaboration with the US Department of Transportation (DOT). The tasks have included near duplicate detection to address mass comment campaigns [58], comment topic modeling [5, 8, 30, 51, 59], stakeholder attitude identification [1, 31], and presence of substantive points in public comments [2, 44, 45, 57]. The RegulationRoom project has generated a number or papers on argument mining and conflict detection within comments [29, 34, 43]. These research efforts have focused on examining only a few regulatory proceedings at a time, whereas we evaluate a signifcantly larger dataset containing hundreds of rule documents.

More recent work on e-regulation has analyzed public comment data collected by `regulations.gov` [13, 14, 35, 37, 50, 52], rule-specific data from the Canadian government [53], and data from the White House e-petition platform [15, 19–21]. The tasks addressed in this body of work are topic modeling [15, 20, 21, 35, 37, 52, 53], sentiment analysis [13, 14, 37, 50], named entity recognition [20], and social network analysis [19].

Segmentation of text into discourse units [38] is a core natural language task. Many downstream tasks, such as information extraction [27], sentiment analysis [3], information retrieval [16], and summarization [4, 36], can benefit from discourse segmentation. Because lexical and syntactic text properties form important discourse clues [6], many segmentation methods rely on hand-crafted features to capture them [17, 26]. Classic learning frameworks that have been used for discourse segmentation are linear Support Vector Machines (SVM) [11] and linear-chain Conditional Random Fields (CRF) [32].

One of the key challenges in discourse segmentation development is the dearth of annotated data, which, until recently, prevented the use of neural architectures. Effective neural discourse segmentation methods [22, 56] have relied on word representations obtained from an external neural model trained to perform a related task using a large corpus [39, 49]. The state-of-the art neural discourse segmentation framework [18, 56] has employed a Bidirectional Long-Short-Term Memory-CRF architecture (BiLSTM-CRF) [25] with an attention mechanism [55].

For our baseline model development, we have combined several classic learning methods with hand-crafted, as well as neural sentence representations, from Bidirectional Encoder Representations from Transformers (BERT) [12], which were trained on English Wikipedia (2,500 million words) and BooksCorpus (800 million words) [60] using masked language and next sentence prediction objectives. BERT representations have demonstrated to perform well on a wide range of natural language processing tasks. We also explore whether fine-tuning of BERT on the unlabeled documents in our corpus improves performance.

3 DATA

3.1 Rule-Making Documents

We work with the EPA's final rule documents that are part of the FR. Along with a summary, each of our documents can contain one or more of the following sections: regulatory background, scope of the regulation, rationale for action, technical material describing the regulatory requirements, responses to public comments on the proposed regulation, statutory and executive order review, and legal references. We are interested in automated identification of all passages where the agency discusses public comments, which could occur throughout the document and are not necessarily confined to the comment response section.

We note that the structure of the final rule documents can vary significantly depending on whether it has been produced by the EPA headquarters or a regional office, as well as depending on the specific EPA office (e.g., Office of Water, Office of Air and Radiation). For example, rule documents produced by the headquarters offices are usually major federal regulations that tend to be long and receive significant public feedback. On the other hand, rule documents produced by regional offices tend to be shorter.[4]

It should be noted that our dataset only contains final rule documents as published in the FR. It does not include submitted comment documents, technical support documents, or detailed, dedicated comment response documents that are part of the docket but extraneous to the register.

3.1.1 Task 1: Detecting Comment Discussions. In the first task, we want to identify the spans in the document where the EPA discusses submitted public comments. Examples of a comment discussion include:

- Descriptions of comments received by the agency. For example, "EPA received comments suggesting that the definition of clean alternative fuel conversion should be limited to a group of fuels with proven emission benefits.";
- Descriptions of the agency's responses to the comments it receives. For example, " EPA believes however that the public interest is better served by a broader definition that allows for future introduction of innovative and as-yet unknown fuel conversion systems. EPA is therefore finalizing the proposed definition of clean alternative fuel conversion...".

By distinction, we are not interested in:

- Summarized feedback from petitions (as opposed to public comments) to the agency;

[4]With the possible exception of the regional air quality rules that still tend to attract considerable public attention

- Descriptions of the public comments on another rule;
- Statements such as "we received no comments";
- Passages discussing revisions of a regulatory standard rather than revisions of the proposed rule;
- Referrals to another document in the docket with detailed responses to comments.

3.1.2 Task 2: Classification of Comment Merit. In the second task, we want to classify each comment discussion span as to whether the discussed comment prompted a change in the final rule from the proposed rule. As such, we are considering three categories: passages in which the agency indicates a revision of the rule based on a public comment, passages in which the agency dismisses a comment, and neutral comment discussion passages (i.e., the passages in which the agency neither dismisses the comment nor indicates a revision).

Examples of formulations reflecting comment-based regulatory change are rule revisions and rule withdrawals:

- "To address concerns about space limitations, EPA will allow the label information to be logically split between two labels that are both placed as close as possible to the original Vehicle Emission Control Information (VECI) or engine label."
- "EPA agrees and is including use of this procedure in the OBD demonstration requirement for intermediate age vehicles."
- "The EPA has reviewed the new data submitted by the commenter and used these data to determine the revised MACT floor for continuous process vents at existing sources."
- "EPA received one adverse comment from a single Commenter on the aforementioned rule. As a result of the comment received, EPA is withdrawing the direct final rule approving the aforementioned changes to the Alabama SIPs."

Examples of comment dismissals without a subsequent regulatory change are:

- "We disagree that our action to approve California's mobile source regulations that have been waived or authorized by the EPA under CAA section 209 is inconsistent with the Ninth Circuit's decision..."
- "EPA is finalizing the conversion manufacturer definition as proposed."
- "While we agree with the commenter that pressure release from a PRD constitutes a violation, we will address this in a separate rulemaking..."
- "In the final rule we will clarify our position..."
- "EPA appreciates support from the commenters for this initiative and agrees that the rule makes it possible for EPA to process the TRI data more quickly."
- "EPA believes that no further response to the comment is necessary..."

We observe that this task requires considerably more complex inference, potentially spanning multiple sections of the document. As seen in the examples above, comment dismissals range from very obvious to rather subtle. In turn, determinations of whether a rule was materially revised based on the public comments may also require a clear understanding of what was proposed in the first place.

An extreme example of this can be seen from the following comment dismissal sentence:

"Certain aspects of good engineering judgment described in the exhaust control system, evaporate control system, and fuel delivery control system sections may be approached differently than described above, but EPA expects that test data demonstrating compliance is required rather than optional in such cases."

The sentence responds to technical objections to a regulation by conceding that alternatives are valid ("may be approached differently") but goes on to state the substantive decision in domain terminology ("compliance is required rather than optional", suggesting that the comment had advocated for the "optional" alternative). Without context, it is unclear whether this sentence has anything to do with comments at all, let alone whether required vs. optional compliance results in it agreeing with, or dismissing, the comment's arguments.

3.2 Acquisition and Sampling

We have created our corpus from `regulations.gov` data by selecting EPA regulatory dockets for rules proposed in January 2003 or later and finalized as of March 2018. Our selection has been constrained to dockets containing at least one proposed rule document, at least one final rule document, and at least one comment document. Our corpus contains 1,566 EPA dockets (meta-data 8.8 MB), 2,645 final rule documents (HTML, 376 MB), 2,531 proposed rule documents (HTML, 400 MB), and 282,655 comment documents (85% PDF, 36 GB; 15% plain text, 836 MB).

For the purposes of exhaustive rule document annotation, we have used stratified random sampling at the docket level to select two development docket sets (dev1 and dev2) and one test docket set. The sampling procedure has ensured that the docket sets are a representative mix of EPA program offices and regions.[5] As such, we have obtained 75 dev1-set dockets (116 documents), 76 dev2-set dockets (136 documents), and 73 test-set dockets (99 documents).

In our qualitative examination of the regulatory documents, we have found that the section headers of the rule documents are often informative about whether a section contains a discussion of public comments. To make use of this additional information, we have applied the same random sampling procedure to the remaining dockets to obtain 211 training dockets (817 training documents) and 103 validation dockets (197 validation documents) for the section header annotation.

3.3 Preprocessing

The rule documents were processed in two steps. First, we have applied a rule-based rule document parsing procedure to delete tables, split the text into sections, and retrieve section titles of the first and second level super-sections. This procedure exploits the regular structure of documents to create heuristics applicable to roughly 90% of documents.[6] When exceptions to the standard structure are detected, we manually fixed irregularities to enable

[5] For example, Office of Water/Headquarters, Office of Air and Radiation/Region 1 – Boston.

[6] For example, the first and the second level sections are numbered consecutively in Roman numbers and Latin letters, respectively.

automatic parsing. Second, the section text has been split into sentences, tokenized, and lemmatized using SpaCy [24][7].

3.4 Annotation

3.4.1 Rule Documents. We hired ten students from Carnegie Mellon University and the University of Pittsburgh to perform the annotation tasks during the period of February 2019–April 2019. All annotators are at least second year undergraduate students. Five of the annotators are masters students in fields including computer science, public health, product management, and international relations. The other five are undergraduate students in civil engineering, creative writing, business, and human computer interaction.

The annotators were trained to perform the two tasks described in Section 3.1.1 and Section 3.1.2. For the first task, each annotator received an hour-long in-person training as well as individualized feedback on a set of four training documents. For the second task, the guidelines were delivered via a video. Each annotator received 50 documents on average, including reliability annotations. The documents were allocated such that each annotator worked on a balanced mix of documents from different EPA offices, regions, and dev1/dev2/test set dockets. The annotations were performed using an online tool developed by a collaborating group at the University of Pittsburgh called *Gloss*.

Finally, we note that some annotators did not complete all assignments for the segmentation task, leading to some redistribution of work. The comment response classification task was completed by eight annotators of the initial ten annotators.

3.4.2 Section Headers. Annotation of the section headers was performed by a sole expert annotator (the first author). To this end, all unique section titles were extracted along with three samples of the first paragraph following the section title. These examples are used to judge whether a section contains comment discussion: If all three sample paragraphs include comment discussions, the section title is flagged as the comment-discussion-indicative title.[8]

4 METHODS

To generate baseline results, we use a classic linear SVM[9] and linear-chain CRF[10] learners to segment the rule documents into spans that contain public comment discussion and merit evaluation by the agency.[11] The benefit of the CRF over the SVM is that, when predicting a sentence label, it takes into account the label of the prior and subsequent sentence in addition to the focal sentence's feature vector. In addition, to understand the impact of incorporating feature interactions, we conduct experiments with the Multi-Layer-Perceptron (MLP)[23].[12]

We estimate three binary sentence-level models predicting whether a given sentence contains: (i) a public comment discussion, (ii) a dismissal of a public comment by the agency, and (iii) an agency decision to revise the proposed rule based on the public comments. For the CRF modeling, a training instance is a sequence of sentences within the rule document section boundaries. The hyperparameters have been tuned by fitting the models to the dev1-set and evaluating results on the dev2-set.

4.1 Handcrafted Features

For sentence representation we concatenate three categories of handcrafted features. First, we featurized the text of the sentence for which the prediction needs to be made, as well as the text of the preceding sentence, and concatenate the feature vectors. We use original tokens (including stop words, but excluding punctuation), modified tokens with attached POS tags, bigrams of modified tokens, and bigrams of POS tags.[13] We apply feature hashing [40] to reduce dimensionality. This results in a feature set of size 2,001.

Second, we featurized the text of the section header containing the sentence in question. In that, we apply the same feature generation process used for sentences to the text of the sentence-bearing section header and the header that precedes it. The dimension of this feature set is 101.

Third, we also add a binary flag equal to one if a header of the section in which the sentence occurs has been predicted to contain a comment discussion. We generate these predictions through instance-based learning on the unique section headers from the training set of dockets set aside for this purpose (see Section 3.4.2). Based on the unique headers from the associated validation docket set, this signal mining procedure has a recall of 0.54 and a precision of 0.88.

4.2 Neural Features

We employ BERT[12] to create embedded vector representations for sentences and section headers. BERT is a state of the art neural network language model trained on a large collection of English text in a quasi-unsupervised fashion by having it learn to predict masked words in a sentence, or to classify whether one sentence follows another, or not. By doing so, BERT learns to maintain a neural representation of language context. These vector representations of English text can then be used as for various natural language processing tasks and have been shown to yield significantly better performance than context-independent word embeddings.

As in case of the hand-crafted features, we concatenate both the vectors of the sentence/header in question as well as the context represented by the preceding sentence/header to form a final feature vector. We explore performance of the available pretrained BERT model as well as a BERT model that has been fine-tuned on approximately 6,000 rule documents from our corpus that have not been included in the annotated document sets. To this end, we rely on a PyTorch[47] implementation of BERT.[14] The size of the

[7] Version 2.0.18 (model en_core_web_sm)

[8] For example, there were several first level section titles "What comments did EPA receive?".

[9] We use scikit-learn version 0.20.2 SVC implementation [48] with an error term penalty parameter of 1, and 1,500 as the maximum number of iterations.

[10] We use PyStruct 0.3.2 implementation [42] of margin re-scaled structural SVM using the 1-slack formulation and cutting plane method [28]. We used regularization parameter of 0.1 and 1,500 as the maximum number of iterations.

[11] We have been unable to fit kernelized polynomial and RBF SVMs to our data because these methods do not scale well to the size of our dataset.

[12] We use a scikit-learn version 0.20.2 MLP implementation [48] with one hidden layer of 100 units optimized for at most 100 epochs at the default settings.

[13] We do not use a TFIDF feature representation because it has not performed as well as a simple count-based featurizer in our preliminary experiments.

[14] PyTorch Pretrained BERT: The Big and Extending Repository of pretrained Transformers from https://github.com/huggingface/pytorch-pretrained-BERT. We used the bert-base-uncased version of the model.

generated sentence/header embedding is 728. The fine-tuned model was trained for seven epochs.

5 EVALUATION

We evaluate the quality of the rule document annotation using Cohen's kappa coefficient [10], as well as qualitatively. Performance of our baseline text segmentation models is evaluated on the test set at the sentence level using area under the ROC curve (AUC), F1-score, precision, and recall. We found a sentence to be the most meaningful operational definition of a passage, because comment-discussing sentences are often interspersed with ignorable sentences of a section or a paragraph. For each model, the classification cutoff has been determined using a threshold that maximizes the F1-score on the training data.

6 RESULTS

6.1 Annotation

Table 1 summarizes the key properties of the annotated dataset. For this summary, we have converted span-level annotations into sentence-level annotations. To this end, we have assigned a label to a sentence if an annotator has marked 80% of tokens that make up that sentence. For documents that have been annotated by multiple individuals, we assign a label to a sentence if at least one individual has labeled the sentence. This approach has been motivated by a qualitative examination of annotations, which revealed low recall issues for some annotators. Depending on the dataset, non-ignorable content (i.e. text labeled as discussing comments) comprises 21% to 33% of all sentences, comment dismissals comprise 4% to 5% of all sentences, and comment-based revisions comprise 2% to 3% of all sentences. Approximately half of all labeled sentences have been annotated by two individuals. Due to the annotator attrition, reliability annotations for a more refined labeling task (i.e., identification of comment dismissals and comment-based rule revisions) are available for 73% to 79% of all double-annotated sentences.

Table 1 also reports the inter-annotator agreement statistics, while Table 2 summarizes agreement with the expert annotator on four final rule documents used as part of the annotator training. (Expert annotations have been produced by the first author, who has 10 years of professional experience in supporting EPA's regulatory proposal development.) For the non-ignorable content, inter-annotator agreement scores range from 0.38 to 0.67 (depending on the dataset), whereas agreement with the expert is 0.74 on average (range: 0.35–0.95). We note that agreement on this task appears to improve from the dev1 set to the test set, which may reflect that the annotators learned to do the task better over time, given the order in which the documents have been assigned. Inter-annotator agreement for the comment dismissal labeling task ranges from 0.18 to 0.32, while agreement on the comment-based rule revisions is very low, ranging between 0.086 and 0.19. Agreement with the expert on these tasks is also low: 0.33 (range: 0–0.54) for the comment dismissals and 0.38 (range: 0–0.75) for the comment-based rule revisions.

We have reviewed the annotator errors vis-a-vis the expert annotator. False negatives tend to occur most commonly when:

- The annotator captures only the initial part of the comment discussion that contains typical lexical cues (e.g., "EPA received comments suggesting...", "Commenters noted...", "EPA agrees with the commenters...") but fails to include the entire—usually technical—comment discussion that can span multiple subsequent paragraphs;
- A passage with comment discussion is "buried" in the middle of a longer paragraph, as often happens when comments are discussed in the background section;
- For the more difficult annotation task of identifying comment-based rule revisions and comment dismissals, we have noted that false negatives tend to occur when the evaluation of the passage requires complex inference. As such, the annotators tended to be conservative about assigning these labels for less obvious examples.

Table 1: Characteristics of the Annotated Data

Characteristic	Dev1-set	Dev2-set	Test-set
Number of the Data Set Elements			
Dockets	75	76	73
Documents	116	136	99
Sections	2,197	2,123	1,766
Sentences	72,969	61,837	61,042
Words	1,820,619	1,583,518	1,430,134
Number of the Annotated Sentences			
Non-ignorable content	19,465	20,105	12,979
Comment dismissals	3,527	3,225	2,202
Comment-based regulatory change	2,092	1,015	1,088
Number of the Double-Annotated Sentences			
Non-ignorable content*	42,296	25,300	41,572
Refined content**	33,595	18,561	32,331
Annotator Agreement (Kappa)			
Non-ignorable sentences*	0.42	0.52	0.67
Non-ignorable sentences**	0.38	0.43	0.64
Neutral comment discussion	0.39	0.44	0.66
Comment dismissals	0.32	0.18	0.29
Comment-based regulatory change	0.086	0.19	0.16
Multi-class	0.33	0.38	0.56

Notes: * Sentences for which double annotation of non-ignorable content is available. ** Sentences for which double annotation of content is also available.

Table 2: Annotator Agreement* with Expert

Kappa	Mean	Min	Max
Non-ignorable content	0.74	0.35	0.95
Comment dismissals**	0.33	0	0.54
Comment-based revisions**	0.38	0	0.75

Notes: * Agreement is calculated at the sentence level for four final rule documents. A total of 4,105 sentences are available for this evaluation. ** These statistics are calculated for the eight annotators who performed the task.

For the false positives, we have observed the following tendencies:

- EPA regulations are typically incremental, in that they often tend to modify older, preexisting rules. Therefore, the final and proposal rule document discuss changes/ revisions of the prior regulatory standard. This has been a significant source of confusion for the annotators, who found it difficult to separate comment-based revisions of the proposed regulation from the revisions of the regulatory standard on the regulatory agenda, leading to false positives.
- Another challenge for the annotators has been the decision of when the discussion switches from comment-related to the general topics, also leading to false positives.
- Specifically for the comment-based rule revisions, some annotators found it challenging to distinguish between revisions of the proposed rule that were based on comments from revisions that occurred for other reasons. For example, the EPA may implement revisions based on new evidence that emerges after the proposed rule is submitted for public review.

6.2 Classification Results

Table 3 shows the test set evaluation performance results for each binary classification task divided by learning framework and feature set. With two exceptions, the models have produced better than random predictions, with largest AUC of 0.928 noted for the non-ignorable content prediction and smallest AUC of 0.677 noted for the comment-based rule change prediction. These patterns largely reflect the differences in the quality of annotations obtained for our prediction tasks, with the segmentation task being significantly easier than the comment response classification task.

For the non-ignorable content prediction, linear models produce recall in the range of 0.435–0.687 and precision in the range of 0.749–0.878. Unsurprisingly, for the more complex annotation tasks with low annotator agreement, classification quality is poor. For the comment dismissal prediction, recall is 0.025–0.445 and precision is 0.194–0.510, whereas for the comment-based rule change prediction, recall is 0.020–0.163 and precision is 0.091–0.175.

6.2.1 Linear Model Analysis. CRF model results do not appear to be materially different from those generated by the SVM model on the same handcrafted feature set, albeit taking into account the labels of neighboring sentences. We note, however, that the CRF models have produced consistently higher recall and F1 scores, compared to the SVM models estimated on the same feature set. Because we experienced some convergence problems with CRF models, we have fit them to only one feature set.

Table 3 also shows that neural BERT features tend to generate higher AUC and higher precision, whereas the handcrafted features yield better recall. We note that for the comment-based revision prediction task with considerable label noise, SVM models based on neural features have collapsed to the majority class predictions. This is likely due to a combination of low quality annotations and sparse labels.

We also observe that neural features based on the fine-tuned BERT can perform better than those using out-of-the-box BERT (e.g. best AUC and precision on non-ignorable content prediction). Interestingly, combining neural and handcrafted feature sets generally does not produce synergy performance increases, which could be due to the substantial increase in the overall feature dimension, or the lack of feature interaction capacity in linear models.

Table 3: Baseline Test Set Results

Model	AUC	F1	Prec.	Recall
All Non-ignorable Content				
SVM+HCF	0.9110	0.7105	0.7941	0.6427
CRF+HCF	n.a.	**0.7172**	0.7498	**0.6873**
SVM+BERT (as is)	0.9210	0.6852	0.7754	0.6138
SVM+HCF+BERT (as is)	0.9146	0.6627	0.8236	0.5543
SVM+BERT (tuned)	0.**9277**	0.5821	**0.8778**	0.4354
SVM+HCF+BERT (tuned)	0.9003	0.6697	0.7494	0.6053
Comment Dismissals				
SVM+HCF	0.7596	0.1665	0.2993	0.1153
CRF+HCF	n.a.	0.2090	0.2254	0.1949
SVM+BERT (as is)	0.8688	0.0725	0.4123	0.0397
SVM+HCF+BERT (as is)	0.8618	**0.2699**	0.1936	**0.445**
SVM+BERT (tuned)	**0.8690**	0.0474	**0.5102**	0.0249
SVM+HCF+BERT (tuned)	0.7498	0.2637	0.2813	0.2481
Comment-based Regulatory Change				
SVM+HCF	0.6773	0.0686	**0.175**	0.0205
CRF+HCF	n.a.	0.0880	0.0911	0.0851
SVM+BERT (as is)	*0.8017*	n.a.	n.a.	n.a.
SVM+HCF+BERT (as is)	0.7355	0.0936	0.0914	0.0958
SVM+BERT (tuned)	*0.8149*	n.a.	n.a.	n.a.
SVM+HCF+BERT (tuned)	**0.7901**	**0.1238**	0.0996	**0.1634**

Notes: HCF – hand crafted features. AUC – area under the ROC curve. CRF model does not produce confidence scores, hence AUC estimation was not possible. The classification cutoff was chosen to maximize F1 score for each model.

6.2.2 Multi-Layer Perceptron Results. In a second set of experiments we assessed whether classification performance increases with models that allow for feature interactions. To this end, we trained a series of Multi-Layer-Perceptron models (i.e. a neural network with one hidden layer of size 100 and a two-class softmaxed output) on our tasks and feature sets. Table 4 contains the results we obtained on two MLP variants that differ by whether they include a Rectified Linear Unit (ReLU) activation function before the final softmax (MLP-ReLU), or an identity transformation (MLP-Id). The practical difference is that a ReLU activation will truncate all incoming negative activation values to 0 and leave positive ones unchanged.

For Task 1, we observe that nonlinear models using BERT features can achieve marginally higher F1 scores than the linear models shown in Table 3 at some cost of AUC. We also see that adding handcrafted features to a model can yield some performance synergy in most conditions. From this we infer that nonlinear models can potentially produce better results on our dataset, and hence we plan to experiment with recurrent or dilated convolutional models

for sequence tagging to leverage the document context in future work.

Linear models perform better than nonlinear models on Task 2 (dismissal and change classification), albeit not well by absolute standards. We attribute this to a combination of larger class imbalance and low quality annotations as evidenced by agreement scores.

Table 4: Auxiliary Test Set Results

Model	AUC	F1	Prec.	Recall
All Non-ignorable Content				
MLP-ReLU+HCF	0.7981	0.6802	0.7189	0.6455
MLP-Id+HCF	0.7957	0.7012	0.8058	0.6206
MLP-ReLU+BERT (as is)	0.7491	0.6381	0.8274	0.5194
MLP-Id+BERT (as is)	0.6872	0.5349	**0.8766**	0.3849
MLP-ReLU+HCF+BERT (as is)	0.8043	0.6709	0.6676	0.6742
MLP-Id+HCF+BERT (as is)	0.8318	**0.7351**	0.7676	0.7053
MLP-ReLU+BERT (tuned)	0.8050	0.6863	0.7120	0.6624
MLP-Id +BERT (tuned)	0.7738	0.6775	0.8363	0.5694
MLP-ReLU+HCF+BERT (tuned)	0.8150	0.6986	0.7147	0.6832
MLP-Id+HCF+BERT (tuned)	**0.8434**	0.7232	0.6984	**0.7499**
Comment Dismissals				
MLP-ReLU+HCF	0.6021	0.2334	0.2376	0.2292
MLP-Id+HCF	0.5855	0.2191	0.2597	0.1895
MLP-ReLU+BERT (as is)	0.5141	0.0555	**0.3974**	0.0298
MLP-Id+BERT (as is)	0.5194	0.0740	0.3562	0.0413
MLP-ReLU+HCF+BERT (as is)	**0.6458**	0.2529	0.2026	**0.3366**
MLP-Id+HCF+BERT (as is)	0.5758	0.2090	0.2819	0.1661
MLP-ReLU+BERT (tuned)	0.5793	0.2129	0.2727	0.1745
MLP-Id +BERT (tuned)	0.6112	**0.2579**	0.2727	0.2447
MLP-ReLU+HCF+BERT (tuned)	0.6021	0.2334	0.2376	0.2292
MLP-Id+HCF+BERT (tuned)	0.5745	0.2100	0.2980	0.1621
Comment-based Regulatory Change				
MLP-ReLU+HCF	0.5276	0.0837	0.1262	0.0626
MLP-Id+HCF	0.5186	0.0661	0.1881	0.0401
MLP-ReLU+BERT (as is)	0.5038	0.0163	0.1059	0.0088
MLP-Id+BERT (as is)	0.5014	0.0058	0.2727	0.0029
MLP-ReLU+HCF+BERT (as is)	0.5273	0.0785	0.1000	0.0646
MLP-Id+HCF+BERT (as is)	0.5207	0.0719	0.1790	0.0450
MLP-ReLU+BERT (tuned)	0.5009	0.0039	0.2857	0.0020
MLP-Id +BERT (tuned)	0.5028	0.0115	**0.3158**	0.0059
MLP-ReLU+HCF+BERT (tuned)	0.5554	0.1168	0.1065	0.1292
MLP-Id+HCF+BERT (tuned)	**0.5573**	**0.1199**	0.1091	**0.1331**

Notes: HCF – hand crafted features. AUC – area under the ROC curve. MLP-ReLU – a multi-layer perceptron with one hidden layer with 100 units and a ReLU non-linearity followed by a Softmax. MLP-Id – a multi-layer perceptron with one hidden layer with 100 units and an identity non-linearity followed by a Softmax; this model is equivalent to a generalized linear regression model with interaction terms. The classification cutoff was chosen to maximize F1 score for each model.

6.2.3 Error Analysis. For our best-performing models we have generated and examined five random examples for each type of error. Our findings are as follows:

False Positives: The models tend to produce false positives when sentences contain certain trigger words (such as "response", "revision", "finalizing the rule as proposed") yet the overall context of the passage is not related to the discussion of public comments. For example, these trigger words have been observed in passages discussing petitions and revisions of the regulatory standard that are not based on comments, similar to mistakes made by human annotators. There is also a fair share of label noise: As noted earlier, the annotators have been challenged by longer comment discussions and occasionally failed to capture the entire relevant span. We also conjecture that in this case the models have been guided by the section-header related signal.

False Negatives: The false negatives tend to occur in sections that do not commonly contain comment discussion (e.g., "Background", "Executive Order Review"). Sentences that lack the boilerplate language (e.g., "response", "EPA", "comment") also tend to be missed more often. As with the false positives, we observed some amount of label noise, often in cases when the annotators mislabeled discussions of regulatory revisions that have not been driven by public feedback or when annotators have failed to determine an appropriate boundaries for the technical discussion of comments.

Label Confusion: We have observed several cases of the models being confused about the polarity of EPA assessment, particularly when the sentence has included trigger words such as "agree" and "disagree" together.

Parsing: We have noted several instances of erroneous sentence parsing (e.g., a citation "40 CFR 51.1010(b)." has been isolated as a sentence) that lead to classification errors. This issue could be remedied by a sentence boundary detector oriented towards processing legal text [54].

7 DISCUSSION

It is likely possible to automatically identify certain type of content in regulatory documents with irregular structure. Our baseline segmentation performance for detecting comment discussion sentences with recall in the range of 0.435–0.687 and precision in the range of 0.749–0.878. While we have focused on identifying comment discussion by the receiving agency, we believe that there are other types of content (e.g., regulatory requirements) automated segmentation of which may be both, desired and feasible. Detecting specific comment discussions that either dismiss comments or announce rule revision turns out to be a harder task for both annotators and, consequently, for models. Moving forward, this begs the question of which information need the model caters to. If value is added by quickly pointing an expert to comment discussion passages, then a well-performing model is within reach given good training data. On the other hand, an automated analysis of topics for which comments have been influential remains a hard problem.

Second, we note that our dataset has been compiled using highly educated non-expert annotators. We have found that this type of

background is sufficient for producing relatively coarse annotations (e.g., identifying parts of the document that contain comment discussion). We have measured the annotator-expert agreement of 0.74 for the comment discussion identification task. However, more refined annotation tasks, such as the ones determining the agency's responses to public feedback, would likely require expert-level understanding of the domain.

We believe that our baseline modeling results can be further improved by developing a fully neural sequence tagging model, such as the one developed for the standard discourse segmentation corpus [56]. However, even with access to the sequence encoders such as BERT, the limited size of our corpus may still present a modeling challenge.

Another immediate options for improvement would be to remedy the label sparsity for the comment dismissal/revision classification by training models only on data that is known to contain comment discussion (i.e. on the non-ignorable sentences) and compose a two tiered model to first detect comment discussions, and then then classify their polarity.

8 CONCLUSIONS

We have produced a dataset and baseline for a novel discourse segmentation task of identifying public comment discussion and evaluation by regulatory agencies. In doing so we presented evidence that detecting comment discussions automatically using mainstream NLP techniques is feasible given good training data. Classifying discussions of a particular type is harder both because of data sparsity and low annotator agreement. While good general detection performance will add value in some practical settings, we see opportunity for further improvement in the use of neural sequence tagging models, albeit subject to the limitations of data quality as a function of annotator expertise, training, and type system design.

9 ACKNOWLEDGMENTS

The authors thank University of Pittsburgh Intelligent Systems Program student Jaromir Savelka for permission to use the *Gloss* annotation tool.

REFERENCES

[1] Jaime Arguello and Jamie Callan. 2007. A bootstrapping approach for identifying stakeholders in public-comment corpora. In *Proceedings of the 8th annual international conference on Digital government research: bridging disciplines & domains*. Digital Government Society of North America, 92–101.

[2] Jaime Arguello, Jamie Callan, and Stuart Shulman. 2008. Recognizing citations in public comments. *Journal of Information Technology & Politics* 5, 1 (2008), 49–71.

[3] Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from rst discourse parsing. *arXiv preprint arXiv:1509.01599* (2015).

[4] Mohammad Hadi Bokaei, Hossein Sameti, and Yang Liu. 2016. Extractive summarization of multi-party meetings through discourse segmentation. *Natural Language Engineering* 22, 1 (2016), 41–72.

[5] Claire Cardie, Cynthia R Farina, Matt Rawding, and Adil Aijaz. 2008. An erulemaking corpus: Identifying substantive issues in public comments. (2008).

[6] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*. Springer, 85–112.

[7] Nuno Carvalho and Rui Pedro Lourenço. 2018. E-Rulemaking: Lessons from the Literature. *International Journal of Technology and Human Interaction (IJTHI)* 14, 2 (2018), 35–53.

[8] Lijun Chen. 2007. Summaritive digest for large document repositories with application to e-rulemaking. (2007).

[9] Cary Coglianese. 2004. E-Rulemaking: Information technology and the regulatory process. *Administrative Law Review* (2004), 353–402.

[10] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[11] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[13] Tao Ding and Shimei Pan. 2016. How Reliable Is Sentiment Analysis? A Multi-domain Empirical Investigation. In *International Conference on Web Information Systems and Technologies*. Springer, 37–57.

[14] Lauren M Dinour and Antoinette Pole. 2017. Potato Chips, Cookies, and Candy Oh My! Public Commentary on Proposed Rules Regulating Competitive Foods. *Health Education & Behavior* 44, 6 (2017), 867–875.

[15] Catherine Dumas, Teresa M Harrison, Loni Hagen, and Xiaoyi Zhao. 2017. What Do the People Think?: E-Petitioning and Policy Decision Making. In *Beyond Bureaucracy*. Springer, 187–207.

[16] Yixing Fan, Jiafeng Guo, Yanyan Lan, Jun Xu, Chengxiang Zhai, and Xueqi Cheng. 2018. Modeling diverse relevance patterns in ad-hoc retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 375–384.

[17] Vanessa Wei Feng and Graeme Hirst. 2014. Two-pass discourse segmentation with pairing and global features. *arXiv preprint arXiv:1407.8215* (2014).

[18] Elisa Ferracane, Titan Page, Junyi Jessy Li, and Katrin Erk. 2019. From News to Medical: Cross-domain Discourse Segmentation. *arXiv preprint arXiv:1904.06682* (2019).

[19] Loni Hagen, Teresa M Harrison, and Catherine L Dumas. 2018. Data Analytics for Policy Informatics: The Case of E-Petitioning. In *Policy Analytics, Modelling, and Informatics*. Springer, 205–224.

[20] Loni Hagen, Teresa M Harrison, Özlem Uzuner, Tim Fake, Dan Lamanna, and Christopher Kotfila. 2015. Introducing textual analysis tools for policy informatics: a case study of e-petitions. In *Proceedings of the 16th annual international conference on digital government research*. ACM, 10–19.

[21] Loni Hagen, Özlem Uzuner, Christopher Kotfila, Teresa M Harrison, and Dan Lamanna. 2015. Understanding Citizens' Direct Policy Suggestions to the Federal Government: A Natural Language Processing and Topic Modeling Approach. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on*. IEEE, 2134–2143.

[22] Mehedi Hasan, A Kotov, S Naar, GL Alexander, and A Idalski Carcone. 2019. Deep neural architectures for discourse segmentation in e-mail based behavioral interventions. In *American Medical Informatics Association (AMIA)*.

[23] Geoffrey E Hinton. 1990. Connectionist learning procedures. In *Machine learning*. Elsevier, 555–610.

[24] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear* (2017).

[25] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* (2015).

[26] Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 13–24.

[27] Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-Level N-ary Relation Extraction with Multiscale Representation Learning. *arXiv preprint arXiv:1904.02347* (2019).

[28] Thorsten Joachims, Thomas Finley, and Chun-Nam John Yu. 2009. Cutting-plane training of structural SVMs. *Machine Learning* 77, 1 (2009), 27–59.

[29] Barbara Konat, John Lawrence, Joonsuk Park, Katarzyna Budzynska, and Chris Reed. 2016. A Corpus of Argument Networks: Using Graph Properties to Analyse Divisive Issues.. In *LREC*.

[30] Namhee Kwon, Stuart W Shulman, and Eduard Hovy. 2006. Multidimensional text analysis for eRulemaking. In *Proceedings of the 2006 international conference on Digital government research*. Digital Government Society of North America, 157–166.

[31] Namhee Kwon, Liang Zhou, Eduard Hovy, and Stuart W Shulman. 2007. Identifying and classifying subjective claims. In *Proceedings of the 8th annual international conference on Digital government research: bridging disciplines & domains*. Digital Government Society of North America, 76–81.

[32] John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).

[33] Gloria T Lau. 2004. *A comparative analysis framework for semi-structured documents, with applications to government regulations*. Stanford University.

[34] John Lawrence, Joonsuk Park, Katarzyna Budzynska, Claire Cardie, Barbara Konat, and Chris Reed. 2017. Using argumentative structure to interpret debates in online deliberative democracy and eRulemaking. *ACM Transactions on Internet Technology (TOIT)* 17, 3 (2017), 25.

[35] Karen EC Levy and Michael Franklin. 2014. Driving regulation: using topic models to examine political contention in the US trucking industry. *Social Science Computer Review* 32, 2 (2014), 182–194.
[36] Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in near-extractive summarization. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 137–147.
[37] Michael A Livermore, Vladimir Eidelman, and Brian Grom. 2017. Computationally assisted regulatory participation. *Notre Dame L. Rev.* 93 (2017), 977.
[38] Daniel Marcu. 2000. *The theory and practice of discourse parsing and summarization*. MIT press.
[39] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
[40] John E. Moody. 1988. Fast Learning in Multi-Resolution Hierarchies. In *Advances in Neural Information Processing Systems 1, [NIPS Conference, Denver, Colorado, USA, 1988]*. 29–39. http://papers.nips.cc/paper/175-fast-learning-in-multi-resolution-hierarchies
[41] Peter Muhlberger, Nick Webb, and Jennifer Stromer-Galley. 2008. The Deliberative E-Rulemaking project (DeER): improving federal agency rulemaking via natural language processing and citizen dialogue. In *Proceedings of the 2008 international conference on Digital government research*. Digital Government Society of North America, 403–404.
[42] Andreas C. Müller and Sven Behnke. 2014. pystruct - Learning Structured Prediction in Python. *Journal of Machine Learning Research* 15 (2014), 2055–2060. http://jmlr.org/papers/v15/mueller14a.html
[43] Joonsuk Park. 2016. *Mining and evaluating argumentative structures in user comments in eRulemaking*. Cornell University.
[44] Joonsuk Park, Cheryl Blake, and Claire Cardie. 2015. Toward machine-assisted participation in eRulemaking: An argumentation model of evaluability. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*. ACM, 206–210.
[45] Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*. 29–38.
[46] Joonsuk Park, Sally Klingel, Claire Cardie, Mary Newhart, Cynthia Farina, and Joan-Josep Vallbé. 2012. Facilitative moderation for online participation in eRulemaking. In *Proceedings of the 13th Annual International Conference on Digital Government Research*. ACM, 173–182.
[47] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS-W*.
[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[49] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).
[50] Rachel A Potter. 2017. More than spam? Lobbying the EPA through public comment campaigns. In *Brookings Series on Regulatory Process and Perspective*. https://www.brookings.edu/research/more-than-spam-lobbying-the-epa-through-public-comment-campaigns
[51] Stephen Purpura, Claire Cardie, and Jesse Simons. 2008. Active learning for e-rulemaking: Public comment categorization. In *Proceedings of the 2008 international conference on Digital government research*. Digital Government Society of North America, 234–243.
[52] Reza Rajabiun. 2015. Beyond Transparency: The Semantics of Rulemaking for an Open Internet. *Ind. LJ Supp.* 91 (2015), 33.
[53] Reza Rajabiun and Catherine Middleton. 2015. Public Interest in the Regulation of Competition: Evidence from Wholesale Internet Access Consultations in Canada. *Journal of Information Policy* 5 (2015), 32–66.
[54] Jaromir Savelka, Vern R Walker, Matthias Grabmair, and Kevin D Ashley. 2017. Sentence boundary detection in adjudicatory decisions in the united states. *Traitement automatique des langues* 58, 2 (2017), 21–45.
[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
[56] Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward Fast and Accurate Neural Discourse Segmentation. *arXiv preprint arXiv:1808.09147* (2018).
[57] Antje Witting. 2015. Measuring the use of knowledge in policy development. *Central European Journal of Public Policy* 9, 2 (2015), 54–62.
[58] Hui Yang and Jamie Callan. 2005. Near-duplicate detection for eRulemaking. In *Proceedings of the 2005 national conference on Digital government research*. Digital Government Society of North America, 78–86.
[59] Hui Yang and Jamie Callan. 2008. Ontology generation for large email collections. In *Proceedings of the 2008 international conference on Digital government research*. Digital Government Society of North America, 254–261.
[60] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*. 19–27.

AI-assisted message processing for the Netherlands National Police

Bas Testerink
Police Lab AI
Netherlands National Police
Driebergen, The Netherlands
bas.testerink@politie.nl

Daphne Odekerken
Police Lab AI
Netherlands National Police
Driebergen, The Netherlands
daphne.odekerken@politie.nl

Floris Bex
Police Lab AI
Utrecht University
Utrecht, The Netherlands
f.j.bex@uu.nl

ABSTRACT

The number of messages that the Netherlands National Police (NNP) receives (e.g. from international partner institutes and citizens) grows steadily every year. The NNP has initiated a number of projects to develop artificial intelligence systems that assist in the processing of such messages. In this demo, we show a prototype of one such system that will be used for supporting the processing of messages from international (Interpol) partners.

CCS CONCEPTS

• **Information systems** → *Expert search.*

KEYWORDS

computational argumentation, decision-support, multi-agent systems

1 INTRODUCTION

The number of messages that the Netherlands National Police (NNP) receives grows steadily every year. Such messages range from notifications from citizens to requests for assistance from international partner institutes. The NNP has initiated a project to develop artificial intelligence (AI) that assists in the processing of such messages, creating autonomous software agents that support human operators. The use of natural language processing tools is a cornerstone of the agents, because incoming messages are typically free-text (e-mails, online forms). Furthermore, it is important that the agents are designed in such a way that every major decision is made transparently, and that legal and ethical rules and regulations can be enforced.

The demo system enhances the existing processing of messages that are received through the Interpol channel. An overview of the goal system incorporated into the Interpol message processing workflow is shown in Figure 1. The pink components with human icons are the human operators. The orange components with the computer chips represent agent components. The yellow components are components without agency.

Currently, a *coordinator* monitors all the incoming messages and categorizes them on priority, theme and relevancy for The Netherlands. The coordinator may answer the message directly, forward the message to a *specialist* for further processing, or choose to ignore the message, for example because it is not relevant for the Netherlands. Specialists specialize in topics such as counter-terrorism and child sex tourism. Usually they have access to domain data and contacts that are relevant for their expertise. A specialist can forward a message internally or answer it directly.

The first agent that is inserted into the workflow is the *Triage Agent*, which supports the coordinator by performing classification and information extraction tasks. What is the theme, priority and relevance of the messages? Which entities (persons, bank accounts, countries, organisations) are mentioned in the message? What is the intent of the sender – are they asking for information, or do they expect the Dutch police to take action? Not all these questions can be answered given just the message. For instance, the occurrence of a person in the police databases may determine the relevancy for The Netherlands. The Triage Agent thus reads the e-mails and supports the coordinator. If the coordinator agrees with the agent, they forward the messages with the relevant annotations (entities, intent, priority level, etc.).

The second type of agent is a *Specialist Agent*, which supports specialists to do routine work on their respective themes. The agents that work for the specialists will formulate the task that is required given the message, execute it, and then report their findings to the specialist as an enhancement of the initial message. The idea is that the specialist ultimately receives messages as if a colleague already processed it. For instance, consider a notification that during a routine border patrol a Dutch vehicle was found to contain illegal drugs. The Triage Agent already determined the vehicle to be indeed Dutch and the message is forwarded to the specialist agent for drug related crime. This specialist agent has access to current drug-related investigations such as which organizations are of special interest. It tries to match the notification to existing investigations or otherwise initiates a new one. By the time the specialist receives the notification from agent he or she can immediately see how the notification relates to past information and what course of action would be prudent. The main task of the specialist then becomes the monitoring and training of the agent.

A final piece of functionality will be to aggregate the messages. At the NNP we are working on real-time monitoring of intelligence data from international partners, combining it with open source intelligence (news, Wikipedia, Twitter) and in-house intelligence from the NNP.

2 AGENT ARCHITECTURE

The architecture we use for the individual (Triage and Specialist) agents is the same architecture that we have used in our other project *Intelligence Amplification for Cybercrime* (IAC) [1], in which we have designed an agent to assist the NNP in the assessment of

In: Proceedings of the First Workshop on AI in the Administrative State, June 17, 2019, Montréal, QC, CA.*AIAS'19, June 17, Montréal, CA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

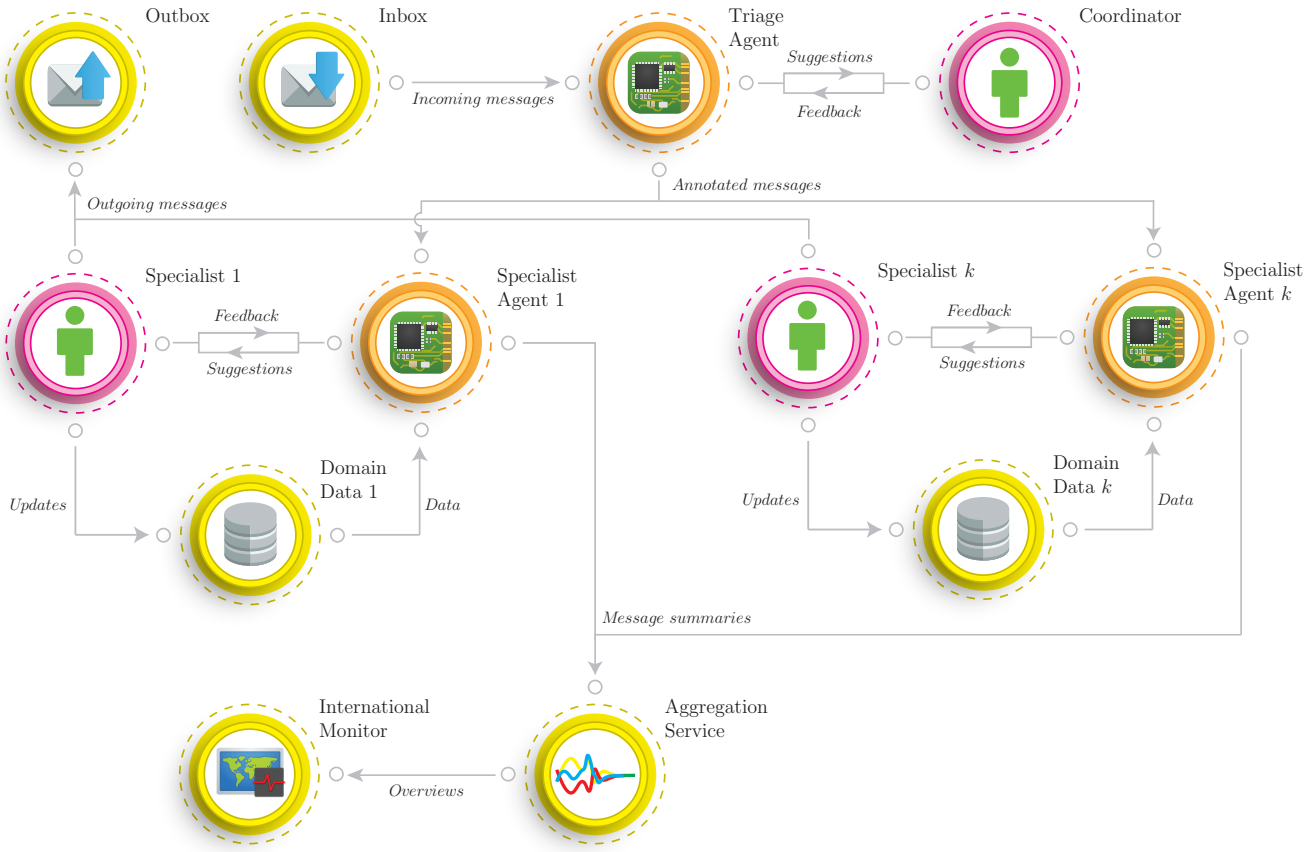

Figure 1: AI-assisted message processing multi-agent system overview.

crime reports submitted by civilians. In a nutshell, the agent applies information extraction techniques to understand a document, applies legal reasoning to determine whether more information is required and applies a policy that is optimized for efficiency to determine the next action.

The agent's goal is to produce some (information) product such as a report, reply or analysis. We refer to a coherent sequence of interactions with the environment as a session. For example, a session can be a sequence of database queries that were required to respond to a message. A session always ends with a terminal action. For instance, terminal actions can be to ignore the message, forward it or answer it directly.

For many law-enforcement applications we need to be able to check why the agent suggests (or directly executes) some terminal action. The basis for many decisions is legislation. Hence, we draw upon the field of computational legal argumentation (cf. [2]) to ensure that the agent has an argument grounded in the relevant rules and regulations when it decides upon a terminal action. The agent architecture is designed for creating agents that efficiently seek information in their environment and transparently decide on what terminal action ought to be executed [4].

Figure 2 shows an overview of the agent architecture. The deployment phase concerns the actual functionality of the agent. The training phase is required to configure the deployment phase components. The deployment components are the top-half (blue) components. The training components are the bottom-half (green) components. The monitor interface and argumentation engine are used in both phases.

2.1 Deployment

The agent connects to its environment through an *external interface*. That interface differs per application. In the message processing system, it will contain functionalities such as forwarding e-mails and querying databases. Typically, the external interface is implemented as a layer that calls different APIs of other systems and passes on the callback. The aim of the agent is to create some (information) product such as a message which is annotated with analyses and suggested actions. These products are stored in the *product database*. Such a product is typically built in two phases: first the agent tries to find enough information to make a final decision on the product, and second the final decision results in a terminal action.

External information, such as a message and database results, are feedback which the external interface sends to the internals of the agent. The feedback is put through a pipeline of *classifiers and attribute extractors* which turn the feedback into structured data (statements about the feedback which are attributes and Boolean

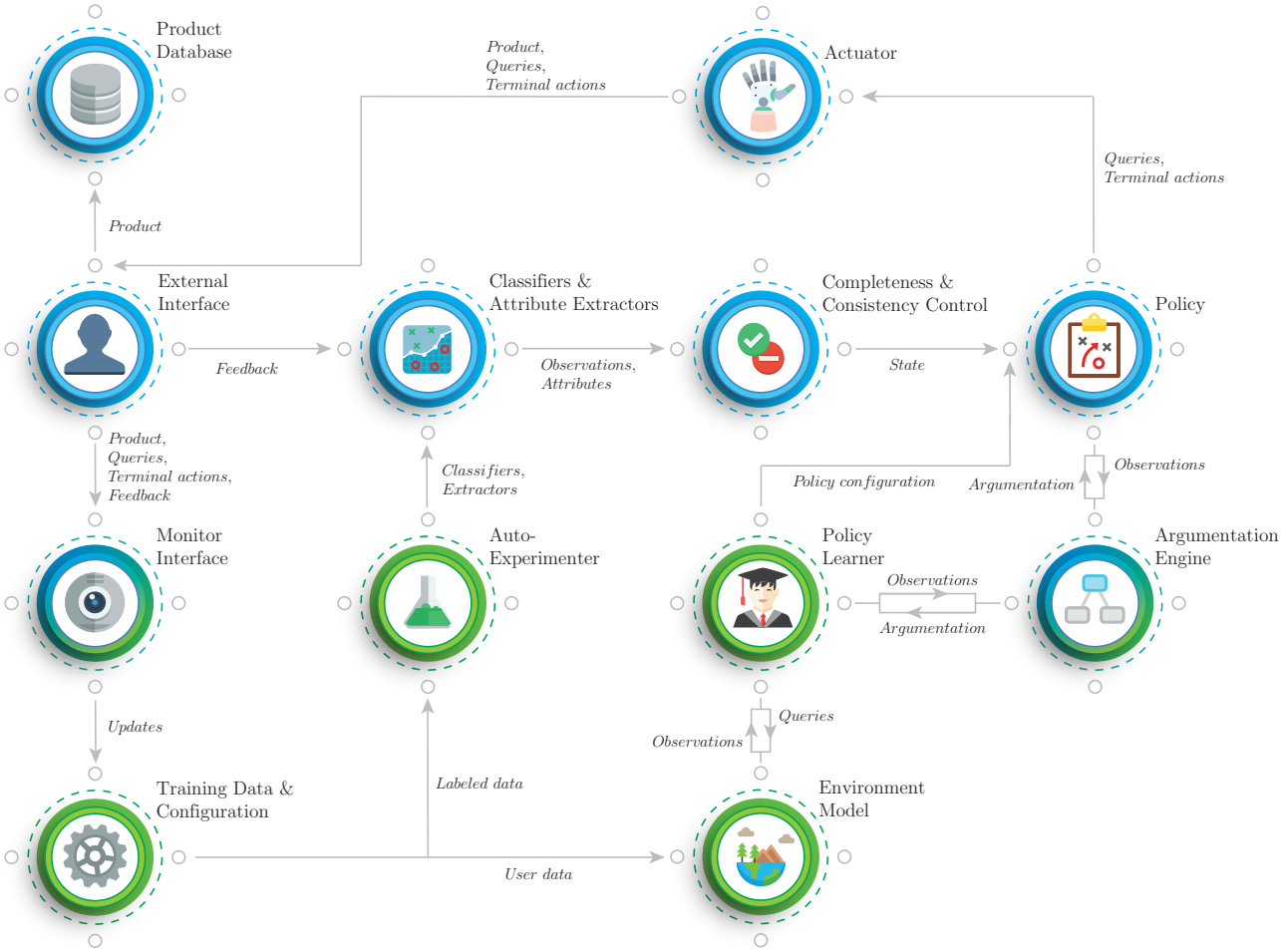

Figure 2: Agent architecture.

observations). For our earlier IAC intake agent, we use existing named entity recognition software [3] and bespoke classifiers (cf. Section 2.2) to classify and extract entities (e.g. the suspect, the victim, addresses) and relations (e.g. "the suspect received money from the victim"). For the Interpol Triage agent we use Spacy as the basis and apply pre- and post-processing to improve upon its base performance. Our choice for Spacy was based on its ease-of-use and available multi-language models for NLP.

The classification and extraction pipeline consists of many separately constructed components which may result in inconsistent results. Hence, we apply a *consistency control* mechanism which makes sure that the data is consistent. It also checks whether the data is complete. That check is mainly for fulfilling the preconditions of final actions, e.g., administrative requirements. The result of this controller is what we consider to be the state of current session, where a session is a sequence of actions after the initial input until the information product is produced (i.e. the suggested course of action for the user).

The actual decision making of the agent is executed by a decision-making *policy*. Based on the state, it determines the next action; this can be an information gathering action (query) or a terminal action. The policy may draw upon the *argumentation engine* in order to argue for or against an action. The *actuator* of the agent prepares the action for execution through the agent's external interface. For instance, the action might be an information gathering action upon a database. The actuator may then formulate an SQL query which the external interface ensures is sent to the appropriate database and returns the result as feedback to the classification and attribute extraction pipeline.

2.2 Training

The *monitor interface* allows a human operator to monitor the agent's activities and control its training phase. The human operator uses the monitor interface also to create labelled data by approving or disapproving (part of) the agent's activities. The monitor interface also shows the argumentation behind core decisions. This connection is not shown in the figure as showing the argumentation does not directly impact the agent's decision-making. However, it does help the human operator to understand the choices of the agent and localize where potential corrections have to be

made. The training of the agent is based on *example data and configuration settings* of its different training tasks. For the observations and attributes, we apply supervised learning which is enabled by the gathering of labelled data during deployment. An *automated experimenter* module tries different algorithms in order to determine for each observation and attribute what the best classifier or extractor is.

The policy of the agent is shaped by reinforcement learning (although other methods can be used). The *policy learner* tries to create a policy which efficiently interacts with the environment. For instance, it may try to minimize the amount of data that it queried from databases. In order to practice these interactions, the policy learner requires an *environment model* that is generated from the training data. The model captures for instance probability distributions over random variables that the agent encounters. At the moment the model's implementation is a Bayesian network where its nodes are observations that the argumentation module may use to construct arguments with. For reinforcement learning, we apply the argumentation engine as part of its reward function. The agent gets a positive reward when it achieves a state such that more feedback from the external interface cannot change its opinion on the terminal action. When such a state is reached, it is natural to opt for the terminal action that the agent can argue for [4].

3 DESIGN CONSIDERATIONS

The application of autonomous A.I. systems requires careful considerations with respect to their potential impact. We decided to restrict the agent's capabilities to reading messages, querying systems and presenting information to the human operator since we cannot guarantee correct behaviour due to the agent's reliance on imperfect information extraction. The agent has no capability for updating databases or sending messages without explicit approval from the human operator.

During deployment, every decision outcome of the agent is documented in its trace for *auditability* and can be inspected by a human operator. However, it should be noted that it is not always possible to completely reproduce the behaviour of the agent: some queried databases contain data that is forbidden by law to store in the same environment. Hence, it is for instance not allowed to store raw database query results. As a result, there are situations in which it cannot be reproduced which information the agent exactly had when it made a decision. This happens for example when source databases are updated. The human operator can provide feedback through the monitor interface which can be taken into consideration when the system is retrained/adjusted.

Interpretability has been an import design influence from the start. It was determined early on that extracting information from data will be a hard to interpret exercise under most circumstances. From this point of view, it was not desirable to design the application as an end-to-end system. Instead, it was opted to create a method where the granularity of extraction can be balanced with interpretability and transparency. In short, we designed the system in such a way that we can choose how much information is obtained through extraction techniques and how much is inferred by argumentation. Generally the trade-off is accuracy vs. transparency.

In order to increase and maintain the *accuracy* of the agent, we rely on three pillars: A) human operators keep providing training data, which is done not only for keeping the data up-to-date, but also to comply with expiry dates of data; B) the auto-experimenter rigorously searches for the best models; and C) collaborations with academia ensure that the latest academic results are tried and tested.

4 CONCLUSION

In this paper we briefly discuss a multi-agent architecture for handling messages from international Interpol partners to the NNP, as well as the architecture of a single agent. In our live demo, we show the workings of a single Triage Agent with a realistic example. We encourage interested programmers to contact the authors for the source code.

REFERENCES

[1] F.J. Bex, J. Peters, and B. Testerink. 2016. AI for online criminal complaints: From natural dialogues to structured scenarios. In *Artificial Intelligence for Justice Workshop (ECAI 2016)*. 22–29.
[2] H. Prakken and G. Sartor. 1996. A dialectical model of assessing conflicting arguments in legal reasoning. In *Logical models of legal argumentation*. Springer, 175–211.
[3] M.P. Schraagen, M.J.S. Brinkhuis, and F.J. Bex. 2017. Evaluation of Named Entity Recognition in Dutch online criminal complaints. *Computational Linguistics in the Netherlands Journal* 7 (2017), 3–16.
[4] Marijn Schraagen, Bas Testerink, Daphne Odekerken, and Floris Bex. 2019 (to appear). Argumentation-driven information extraction for online crime reports.

A Virtuous Circle
Artificial Intelligence and Accessibility for Administrative Applications

Sara Frug
Legal Information Institute
Cornell University
Ithaca NY United States
sara@liicornell.org

Thomas Bruce
Legal Information Institute
Cornell University
Ithaca NY United States
tom@liicornell.org

ABSTRACT
In this position paper, we suggest that accessibility is an emerging, underfulfilled legal requirement that presents not only a potential locus for activity but also an avenue for research. We describe the use of machine-learning-based image classification as a managerial support tool for accessibility enhancement, and suggest directions for further research. Although this discussion focuses on the government information landscape in the United States, the adoption of the Web Content Accessibility Guidelines in the European Union extends its applicability.

In: Proceedings of the First Workshop on AI in the Administrative State, June 17, 2019, Montreal, QC, CA.

CCS CONCEPTS
• Accessibility • Assistive technologies • People with disabilities

KEYWORDS
Accessibility, Artificial Intelligence, Regulations

1 Information Accessibility and Government Administration

The availability of government information is well accepted as a requirement for efficient public administration. Machine-readability of administrative information, although frequently acknowledged as a goal, is often neglected. As a basis for accessibility for the disabled, it receives even less attention. This discussion focuses on Web Accessibility, although it views web accessibility as a consequence of document accessibility. Although this discussion focuses on the United States, the adoption of the Web Content Accessibility Guidelines in the European Union [1] extends its applicability.

1.1 Regulatory Requirements

In the United States, the 1998 amendments [2] to The Rehabilitation Act of 1973 [3] explicitly require that federal electronic and information technology (EIT) be accessible to people with disabilities. The regulations promulgated under the 1998 amendments required adoption of standards consistent with (but

not identical to) the Web Content Accessibility Guidelines Version 1.0 Level A. [4] In 2017, the regulations were refreshed to incorporate by reference the Web Content Accessibility Guidelines Version 2.0.[5]

1.2 Document Accessibility and Web Accessibility Content Guidelines

The Web Content Accessibility Guidelines provide both specific requirements and a general framework for understanding what makes a document accessible. The acronym "POUR" (Perceivable, Operable, Understandable, Robust) summarizes these requirements, the most fundamental of which ensure that information (e.g., words) not be locked in a medium (e.g., a picture PDF) that cannot be perceived by a person with a disability (e.g., blindness). [6]

1.3 Non-Compliance

In 2008 (ten years after the 1998 amendments), the Digital Communications Division of the Department of Health and Human Services (HHS) wrote:

> "Section 508 requires that Web sites and associated content created with federal funding, whether internal or external, government- or contractor-hosted, are accessible to persons with disabilities. The law has been in effect since June 21, 2001. Federal compliance – including that of HHS -- has lagged." [7]

By that point, the 2.0 version of the Web Content Accessibility Guidelines was about to be released. HHS's compliance timetable put completion at 2013.

In 2018, WCAG 2.0 became the standard for Federal websites. The safe harbor provision, however, protected legacy content.

> "This safe harbor provision applies on an "element-by-element" basis in that each component or portion of existing ICT is assessed separately. In specifying "components or portions" of existing ICT, the safe harbor provision independently exempts those aspects of ICT that comply with the existing 508 Standards from mandatory upgrade or modification after the final rule takes effect. This means, for example, that if two paragraphs of text are changed on an agency Web page, only the altered paragraphs are required to comply with the Revised 508 Standards; the rest of the Web page can remain "as is" so long as otherwise compliant with the existing 508 Standards." [5]

As of this writing, even Section508.gov and 18F's Accessibility Guide yielded accessibility errors.

Beyond the protection of the safe harbor, government agencies persist in publishing new, non-accessible content. Most prominently, on April 18, 2019, the U.S. Department of Justice released the much-anticipated so-called Mueller Report as an image-PDF, downloadable from a web page that displayed the following notice:

> "The Department recognizes that these documents may not yet be in an accessible format. If you have a disability and the format of any material on the site interferes with your ability to access some information, please email the Department of Justice webmaster. To enable us to respond in a manner that will be of most help to you, please indicate the nature of the accessibility problem, your preferred format (electronic format (ASCII, etc.), standard print, large print, etc.), the web address of the requested material, and your full contact information, so we can reach you if questions arise while fulfilling your request." [8]

Although the most high-profile, this is far from the only example of new, non-compliant content published on federal agency websites.

1.4 Publication Practices

The Mueller Report is a good example of a general data impoverishment phenomenon in government publishing, which deserves to be the object of attention from all communities that consume government information. The Mueller Report could not have been drafted as a set of pictures of words; rather, the original, machine-readable document had to have been converted for publication into a set of pictures. This data-impoverishment process is not unique to this document - it can be observed throughout the Code of Federal Regulations. Documents that had to have been authored electronically are converted to pictures for publication, leaving the data consumers to "unscramble the egg" and convert them back into machine-readable data formats.

2 Artificial Intelligence and Document Accessibility

Although there is promising work, notably from Rohatgi [9], Wu et al. [10], and Choi et al. [11], to support extraction of machine-readable data from images of charts, graphs, and other data artifacts, for researchers and application developers, common image types have not been addressed systematically.

2.1 Workflows, Experimentation, and Decision Support

LII has begun a pilot project to establish a data conversion workflow and support automation efforts for data-de-impoverishment. The approach has been three-pronged: 1) manually sort and convert figures to SVG and images of equations to MML; 2) annotate SVG images with descriptions of their content; 3) research machine-readable data sources represented as pictures; 4) apply machine-learning techniques to provide decision support for human annotation and conversion.

The pilot project involved collaboration from a specialist in graphics conversion, law and computer science students, and LII's text specialist. The graphics conversion specialist analyzed 14,486 images from the Code of Federal Regulations and sorted them into categories, such as math (6255), diagrams (1410), data tables (1238), maps (3194), forms (1892), labels (351) and logos (77) (some outlier categories, such as photographs, were discovered in the process). Images transformed prior to this project (1149) were sorted into math (241) and non-equations (908) and set aside for testing. The images were grouped according to which areas of the CFR they appeared in and prioritized according to how much web traffic each containing document (section or appendix) received on the LII website. As of this writing, the graphics conversion specialist has converted 2913 math elements to MML and 1005 diagrams to SVG format. Also as of this writing, law students have located alternate sources for 2706 images, most notably over 90 images of pages from the 1991 Standards for Accessible Design as Originally Published on July 26, 1991. The data that has been gathered and generated in this process will be reusable for other such endeavors.

In the process of planning our accessibility project, we discovered the following. First, manual annotation of images has proceeded quite slowly compared with other tasks. As of this writing, fewer than 100 image annotations have been completed. Second, math conversion is much faster than SVG conversion. Third,

sorting for the purposes of identifying good candidates for SVG conversion produces a different categorization than sorting for purposes of distinguishing similar content.

Because we wished to deploy newly-accessible content as quickly as possible, we focused on techniques that would enable us to quickly prepopulate a queue with mathematical content, which is easy both to classify and convert. At the same time, the classification process provides additional clues to aid in re-sorting non-mathematical images for further treatment. Using Keras and OpenCV, we trained a classifier on the eCFR images for the purpose of identifying math. Initial results yielded precision 0.86 and recall 0.88. In practical terms, this approach immediately identified 215 out of 243 math images for conversion and incorrectly identified only 35 out of 875 non-math images. This enables us to speed deployment by prepopulating a queue through automation.

FUTURE WORK

The initial proof-of-concept effort simplified the task to address identification of mathematical images and non-mathematical images. This pre-sorting is adequate for estimation purposes and enables us to generate machine-readable data before comprehensive sorting is complete.

Because conversion projects frequently include tabular data, forms, and textual images, training the model using additional categories would be quite valuable. Because images may contain mixed content, feature identification and multi-label classification are natural areas for further work.

The initial proof-of-concept effort deliberately eschewed image preprocessing. Characteristics of the images suggest techniques for producing more robust and comprehensive models. For example, basic case-insensitive extraction detected image labels - variants of the terms “figure” (1395), “illustration” (19), “plate” (240), or “legend” (410) - in approximately 14% of the training-set images. Because the choice to annotate within the image rather than within the text surrounding the image should be arbitrary, and because images classified as equations almost never have a legend, it seems worthwhile to purge the image legend before training.

Finally, thus far, we have not take advantages of metadata external to the images themselves. Because the images in question are embedded within documents that are published on the web, several additional variables could be made available to the model. The training data could include the catchline for the section or appendix within which the image appears; the full structural location of that document; the text, if any, immediately preceding or following the image; terms assigned to the containing document from an unsupervised topic model; terms assigned to the containing Part by the Office of the Federal Register; even variables such as co-location within a single document or volume of web-traffic to the containing document could prove relevant to image type and could be worth testing.

CAVEATS AND CONCLUSIONS

As mentioned earlier, in this pilot study, the greatest impediment to training a model proved to be some subtle and some not-so-subtle differences between the type of classification needed to support professional workflow and the type of classification that would support automated extraction. Because our preferences for populating the workflow in this instance were determined by the volume of traffic

and co-location of images within a section, several types of content were not distinguished in the initial sorting. For example, where multi-page forms appeared, images containing entirely textual content (such as full pages of instructions) were not distinguished from the form pages for which they provided guidance. In order to produce useful decision-support tools,

Law-and-AI researchers who work on public administration should be aware that the Access Board estimated day-forward web-accessibility compliance resources for the federal government at 5% of web development, software development, and audio-visual production costs, plus an additional 1.25% for evaluation. Should comprehensive conformance become a requirement, the costs will increase accordingly. The Office for Civil Rights of the U.S. Department of Education has, of late, included web accessibility in its enforcement of Section 504 of the Rehabilitation Act, which requires comprehensive equal access to educational services for recipients of federal funding; this means that, as a rule, universities are scrambling to bring their websites into conformance with WCAG 2.0 level AA. [12] Finally, the number of ADA lawsuits treating websites as public accommodations has increased dramatically during the past few years. [13] Reducing data impoverishment in the publication process should limit the need for such work to addressing the challenge of converting non-born-digital images. The combination of labor required and urgency of need makes AI-enhanced automation a timely and valuable avenue for research. An increased focus on document accessibility can create a virtuous circle in which artificial intelligence applications will both help create, and benefit from, the availability of more machine-readable data.

ACKNOWLEDGMENTS
Our thanks to the LII development team, Sylvia Kwakye, Nic Ceynowa, Ayham Boucher, and Jim Phillips and to Point.B Studios.

REFERENCES
[1] Directive (EU) 2016/2102 of the European Parliament and of the Council of 26 October 2016 on the accessibility of the websites and mobile applications of public sector bodies. ELI: http://data.europa.eu/eli/dir/2016/2102/oj.
[2] Electronic and information technology. 29 U.S.C. § 794d. Retrieved from https://www.law.cornell.edu/uscode/text/29/794d.
[3] Pub.L. 93–112, 87 Stat. 355, enacted September 26, 1973), codified as 29 U.S.C. § 701 et seq. https://www.govinfo.gov/content/pkg/STATUTE-87/pdf/STATUTE-87-Pg355.pdf.
[4] Architectural and Transportation Barriers Compliance Board. Electronic and Information Technology Accessibility Standards. 2000. https://www.federalregister.gov/documents/2000/12/21/00-32017/electronic-and-information-technology-accessibility-standards .
[5] Architectural and Transportation Barriers Compliance Board. Information and Communication Technology (ICT) Standards and Guidelines. (Final Rule). 2017. 82 FR 5790. https://www.federalregister.gov/documents/2017/01/18/2017-00395/information-and-communication-technology-ict-standards-and-guidelines.
[6] W3C. Web Content Accessibility Guidelines (WCAG) 2.0. 2008. https://www.w3.org/TR/WCAG20/#intro-layers-guidance.
[7] United States Department of Health and Human Services. 508 Web Compliance and Remediation Framework. 2008. Retrieved by the Internet Archive on 2/6/2018.

https://web.archive.org/web/20180206161308/https://www.hhs.gov/web/section-508/compliance-and-remediation/framework/index.html .
[8] Special Counsel's Office. Report on the Investigation into Russian Interference in the 2016 Presidential Election. 2019. https://www.justice.gov/storage/report.pdf.
[9] Ankit Rohatgi. WebPlotDigitizer. Version 4.2. 2019. https://automeris.io/WebPlotDigitizer.
[10] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17). ACM, New York, NY, USA, 1180-1192. DOI: https://doi.org/10.1145/2998181.2998364.
[11] J. Choi, S. Jung, D.G. Park, J. Choo, and N Elmqvist. 2019. Visualizing for the Non-Visual: Enabling the Visually Impaired
to Use Visualization. Eurographics Conference on Visualization (EuroVis) 2019, Computer Graphics Forum, Vol. 38, No. 3. http://users.umiacs.umd.edu/~elm/projects/vis4nonvisual/vis4nonvisual.pdf .
[12] Lindsay McKenzie, Feds Prod Universities to Address Website Accessibility Complaints. 11/16/2018. Inside Higher Education. https://www.insidehighered.com/news/2018/11/06/universities-still-struggle-make-websites-accessible-all .
[13] Lindsay McKenzie, 50 Colleges Hit With ADA Lawsuits. 12/10/2018. https://www.insidehighered.com/news/2018/12/10/fifty-colleges-sued-barrage-ada-lawsuits-over-web-accessibility .

Towards Measuring Risk Factors in Privacy Policies

Najmeh Mousavi Nejad*
Fraunhofer IAIS & University of Bonn
Sankt Agustin, Germany
nejad@cs.uni-bonn.de

Damien Graux
Fraunhofer IAIS
Sankt Agustin, Germany
damien.graux@iais.fraunhofer.de

Diego Collarana
Fraunhofer IAIS
Sankt Agustin, Germany
diego.collarana.vargas@iais.fraunhofer.de

ABSTRACT

The ubiquitous availability of online services and mobile apps results in a rapid proliferation of contractual agreements in the form of privacy policies. Despite the importance of such consent forms, the majority of users tend to ignore them due to their content length and complexity. Thus, users might be consenting policies that are not aligned to regulations in laws such as the GDPR from the EU law. In this study, we propose a hybrid approach which measures a privacy policy's risk factor applying both supervised deep learning and rule-based information extraction. Benefiting from an annotated dataset of 115 privacy policies, a deep learning component is first able to predict high-level categories for each paragraph. Then, a rule-based module extracts pre-defined attributes and their values, based on high-level classes. Finally, a privacy policy's risk factor is computed based on these attribute values.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; *Neural networks*; • **Security and privacy** → *Privacy protections.*

KEYWORDS

Privacy policy, Deep learning, Rule-based information extraction, Risk factor

1 INTRODUCTION

In the current digital era, almost everyone is exposed to accepting contractual agreements in the form of privacy policies. However, the majority of people skip privacy policies due to their length and complex terminology. According to a recent survey, from 543 university students, only 26% did not choose the 'quick join' routine, while joining a factious social network and unsurprisingly, their average reading time was only 73 seconds [2]. Moreover, for the administrative state is it important to validate the compliance the privacy policies with a correspondent law. For example, the EU regulation General Data Protection Regulation (GDPR) states that the retention period must be specified and limited.

To assist end-users with consciously agreeing to the conditions, we can apply Natural Language Processing (NLP) and Information Extraction (IE) to present a privacy policy in a structured view. Our approach applies supervised deep learning using an annotated dataset (named OPP-115), to assign high-level classes to a privacy policy's paragraphs. Then, according to predicted classes, we define hand-coded rules based on experts annotations, to extract attributes values from each paragraph. Finally, having detailed information for each paragraph, a risk measurement function computes a risk factor based on extracted information. Consequently, a user could choose to stop using a website, if the predicted risk score is high. Additionally, this structured view can be also used by the administrative state to perform a shallow compliance checking.

OPP-115 is a widely-used dataset in the context of privacy policy analysis [5]. It contains in-depth annotations for 115 privacy policies at paragraph level and each paragraph was annotated by 3 experts. There are two types of annotations: high-level classes which define 10 data practice categories; and low-level attributes which include mandatory and optional attributes. For instance, the high-level class *First Party Collection/Use* has 3 attributes: *Collection Mode (explicit or implicit)*, *Information Type (financial, health, contact, location, etc.)* and *Purpose (advertising, marketing, analytics, legal requirement, etc.)*.

The approach proposed in this paper, is built upon on our previous effort, which exploits OPP-115 and deep learning to solve a multi-label classification problem. We feed privacy policy's paragraphs along with the predicted classes into a rule-based IE component and retrieve attribute values. The rules are defined based on OPP-115 low-level annotations. Finally, all predicted categories and extracted information are passed into a risk measurement module and a risk factor will be computed based on hand-coded rules.

The paper is divided into the following sections: in Section 2, we provide an overview of existing effort on measuring risks in privacy policies; Section 3 presents our proposed approach and our evaluation scheme; and finally Section 4 will conclude this paper.

2 RELATED WORK

In light of the, now enforced EU-wide, General Data Protection Regulation (GDPR) [4], there has been an increasing interest towards privacy policy analysis as this new set of regulations increases the constrains for companies holding customers data. Here, we provide a brief overview of studies that specifically addressed risk levels in privacy policies.

`Polisis` is an online service for automatic analysis of privacy policies [1]. Along with classification and structured presentations of privacy policies, it assigns privacy icons which are based on the *Disconnect*[1] icons. These icons include *Expected Use*, *Expected Collection*, *Precise Location*, *Data Retention* and *Children Privacy*. For instance, *Data Retention* color assignments are: Green for retention periods of less than a year; Yellow, when the retention period is longer than one year; and Red, when there is no data retention

In: Proceedings of the First Workshop on AI in the Administrative State, June 17, 2019, Montréal, QC, CA.

[1] https://disconnect.me/

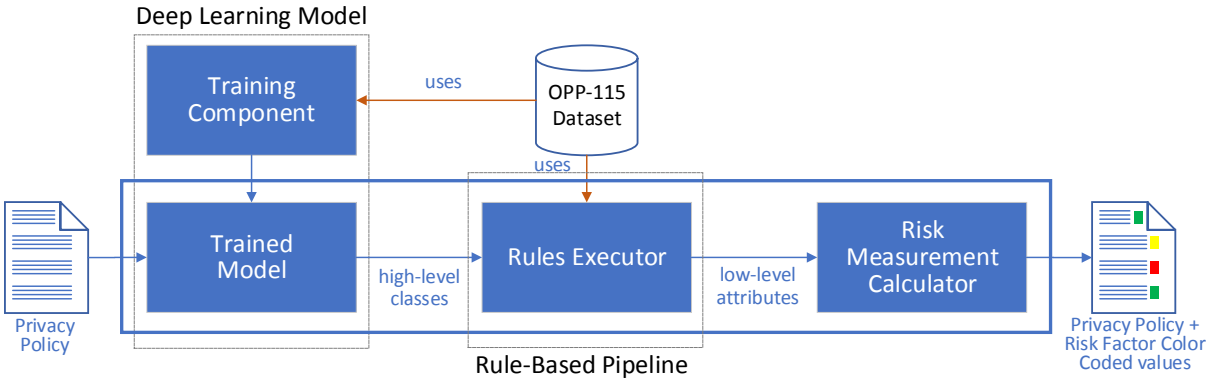

Figure 1: General Architecture.

policy provided. `Polisis` benefits from OPP-115 and employs supervised machine learning to extract high-level categories (in the above example, *Data Retention*) and attribute values of each category (e.g., *Retention Period* in this case). Finally, based on retrieved attribute values and heuristic rules, privacy icons along with their colors are produced. Currently, `Polisis`'s interface generates only a limited set of privacy icons. In future, we intend to further analyze privacy icons and extend them with the help of legal experts.

`PrivacyCheck` is an approach for automatic summarization of privacy policies using data mining [6]. It answers 10 pre-defined questions concerning privacy and security of users' data and is also available as a Chrome browser extension. In order to train the model, a corpus containing 400 privacy policies was compiled and 7 privacy experts manually assigned risk levels (Green, Yellow, Red) to the 10 factors. First, a pre-processing step finds those paragraphs that have at least one keyword related to one of 10 factors. The methodology of selecting keywords was largely manual. Then, the selected paragraphs will be sent to a data mining server where 11 data mining models were trained, one for checking if the corresponding page is a privacy policy and one each for the 10 questions. The authors claim that on average, 60% of the times, `PrivacyCheck` finds the correct risk level. The limitation of `PrivacyCheck` is its lack of Inter Annotator Agreement (IAA) for the annotators. According to the paper, the quality control was performed by assigning each policy to two team members. However, only 15% of privacy policies were compared and their discrepancies were resolved which makes the training dataset less reliable.

`PrivacyGuide` is another summarization tool inspired by GDPR that classifies a privacy policy into 11 categories using NLP and machine learning and further measures the associated risk level of each class [3]. Similar to previous studies, `PrivacyGuide` uses the three-level scale risk based on classification (i.e. Green, Yellow, Red). The 11 criteria and their associated risk levels were defined by GDPR experts. Based on these criteria, a privacy corpus was compiled with the help of 35 university students. Each participant assigned a privacy category to text snippets and classified them with a risk level. The author reported that the weighted average accuracy is 74% for classifying a privacy policy into one of the 11 classes and the accuracy of risk level detection is 90%. Although the results were encouraging, the dataset was not annotated by experts which is a fundamental criterion in legal text processing and analysis.

3 PROPOSED APPROACH

In this section, we provide details of our approach for measuring a privacy policy's risk factor. Our proposed method leverages OPP-115 annotated dataset for training and evaluation [5]. As discussed earlier, OPP-115 high-level annotations are divided into 10 classes:

(1) *First Party Collection/Use*: how and why the information is collected.
(2) *Third Party Sharing/Collection*: how the information may be used or collected by third parties.
(3) *User Choice/Control*: choices and controls available to to users.
(4) *User Access/Edit/Deletion*: if users can modify their information and how.
(5) *Data Retention*: how long the information is stored.
(6) *Data Security*: how is users' data secured.
(7) *Policy Change*: if the service provider will change their policy and how the users are informed.
(8) *Do Not Track*: if and how Do Not Track signals[2] is honored.
(9) *International/Specific Audiences*: practices that target a specific group of users (e.g., children, Europeans, etc.)
(10) *Other*: additional practices not covered by the other categories.

In addition, each high-level category includes low-level attribute annotations. For instance, *Data Retention* category is further annotated with its attributes, which are: *Retention Period*, *Retention Purpose* and *Information Type*. The annotators provided either one or several values for each attribute along with the span of text based on which they have chosen that specific value(s). In the above example, *Retention Period* may have one of the following values: *stated period*, *limited*, *indefinitely* or *unspecified*.

Figure 1 shows the architecture of our proposed approach which consists of three main components: 1) a deep learning module is trained to predict high-level classes of a policy's paragraphs; 2) a rule-based pipeline in which the rules are defined based on low-level attribute annotations of OPP-115; and 3) a risk measurement function that assigns risk icons along with their corresponding colors (green, yellow, red), according to extracted information.

[2]https://en.wikipedia.org/wiki/Do_Not_Track

Table 1: Sample rules for extracting values of Retention Period from Data Retention Category.

Rule	Value	Sample
`[delete/remove][Token]*[after][number][day/month/year]`	Stated Period	1. We remove the entirety of the IP address after 6 months. 2. All stored IP addresses, except the account creation IP address, are deleted after 90 days.
`[not][Token]*[delete/remove]`	Indefinitely	The posts and content you made will not be automatically deleted as part of the account removal process.
`[store/keep/retain/maintain][Token]*[indefinitely]`	Indefinitely	1.This data is generally retained indefinitely. 2. The information we collect for statistical analysis and technical improvements is maintained indefinitely.
`[store/keep/retain/maintain][Token]*[as long as][Token]+`	Limited	1. We will retain your information for as long as your account is active or as needed to provide you services. 2. We will retain your personal information while you have an account and thereafter for as long as we need it for purposes not prohibited by applicable laws
`If not one of the above conditions`	Unspecified	1. We receive and store certain types of information whenever you interact with us. 2. The personal information collected about you through our online applications and in our communications with you is stored in our internal database.

Following conventional ML practices, in the deep learning component, dataset splits are randomly partitioned into a ratio of 3:1:1 for training, validation and testing respectively; while maintaining a stratified set of labels. We further decomposed the *Other* category into its attributes: *Introductory/Generic*, *Privacy Contact Information* and *Practice Not Covered*. Therefore, considering that a paragraph in the dataset may be labeled with more than one category, we face a multi-label classification problem with 12 classes. The implementation of the ML component is completed and we achieve 79% micro-average for F1.

The high-level predicted classes are passed to the rule-based component where low-level attribute values will be extracted. The definition of rules are based on experts annotations in OPP-115 dataset. We intend to use 60% of low-level annotations for defining the rules, 20% for validating the defined rules and the remaining 20% for the final test. Table 1 shows some sample rules for finding values of *Retention Period* attribute in *Data Retention* category. We found our rules definitions based on experts annotations. As shown in the table, the rules definition use the knowledge about high-level categories predicted by the deep learning component.

Algorithm 1 Sketch of risk measurement algorithm

Require: predicted high-level category, extracted attribute values
for all paragraphs in the privacy policy **do**
  $category \leftarrow$ predicted high-level category
  if $category \in$ *Data Retention* **then**
    $RetentionPeriod \leftarrow$ extracted retention period
    if $RetentionPeriod \in$ *(Stated Period, Limited)* **then**
      $DataRetentionIcon \leftarrow$ Green
    else if $RetentionPeriod \in$ *Indefinitely* **then**
      $DataRetentionIcon \leftarrow$ Yellow
    else
      $DataRetentionIcon \leftarrow$ Red
    end if
  end if
  if $category \in$ *First Party Collection/Use* **then** ...
  end if
end for
Ensure: risk icons and their corresponding colors

Having information about attribute values, the risk measurement module is able to assign appropriate risk icons along with their corresponding colors. As a proof-of-concept, we will found our risk measurement rules on *Disconnect* icons. Aforementioned in literature review, the *Disconnect Data Retention* color assignment are as follows: Green for retention period <= 12 months; Yellow, for retention period > 12 months; and Red, when there is no data retention policy provided. Algorithm 1 shows our interpretation of *Data Retention* icon. It is worth to mention that our interpretation is based on the available annotations from OPP-115 dataset. Hence, it is not the only representation that can be built from *Disconnect* icons and others may adopt their own understanding.

For the evaluation of our approach, we intend to generate risk factors according to OPP-115 experts annotations and use it as a goldstandard. We believe the final error will be close to sum of error rate in the deep learning module (predicting high-level classes) and the error which is caused due to incomplete set of rules in rule executor component. Considering the fact that we are now able to predict the correct high-level classes with 79% F1, with the careful definition of rules for extracting attribute values, it is predicted to gain a reasonable accuracy at the end of our pipeline.

4 CONCLUSION

In this study, we proposed the application of Deep Learning models and Rule-Based Information Extraction to automatically present a structured view of risk factors in privacy policies. In particular, we presented a hybrid approach that takes advantage of the dataset OPP-115. This approach is of paramount importance to support users to consciously agree with terms and conditions of online services, and to perform shallow compliance checking where a high-risk score can be assigned to "indefinitely" and "unspecified" values. As next steps, we plan to implement the proposed architecture and run empirical evaluations to validate the presented hypothesis, i.e, users will be more motivated to read privacy policies when a color-coded structured view is presented to them.

REFERENCES

[1] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K. G. Shin, and K. Aberer. Polisis: Automated analysis and presentation of privacy policies using deep learning. *CoRR*, abs/1802.02561, 2018.
[2] J. A. Obar and A. Oeldorf-Hirsch. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, pages 1–20, 2018.
[3] W. B. Tesfay, P. Hofmann, T. Nakamura, S. Kiyomoto, and J. Serna. Privacyguide: Towards an implementation of the eu gdpr on internet privacy policy evaluation. In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*, IWSPA '18, pages 15–21, New York, NY, USA, 2018. ACM.
[4] P. Voigt and A. Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.
[5] S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala, P. G. Leon, M. S. Andersen, S. Zimmeck, K. M. Sathyendra, N. C. Russell, et al. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1330–1340, 2016.
[6] R. N. Zaeem, R. L. German, and K. S. Barber. Privacycheck: Automatic summarization of privacy policies using data mining. *ACM Trans. Internet Technol.*, 18(4):53:1–53:18, Aug. 2018.

Automated Directive Extraction from Policy Texts

Karl Branting
Jim Finegan
David Shin
Stacy Petersen
The MITRE Corporation
McLean, VA, USA
lbranting,jfinegan,hshin,spetersen@mitre.org

Carlos Balhana
Language Technology Lab
University of Cambridge
Cambridge, UK
ceb81@cam.ac.uk

Alex Lyte
The MITRE Corporation
Bedford, MA, USA
alyte@mitre.org

Craig Pfeifer
The MITRE Corporation
Ann Arbor, MI, USA
cpfeifer@mitre.org

ABSTRACT

Federal agencies must comply with directives expressed in documents issued by authoritative sources elsewhere in the government. To automate identification of these directives, the ADEPT (Automated Directive Extraction from Policy Texts) system exploits the insight that directive sentences are usually characterized by deontic modality (e.g. "must", "shall", etc.) permitting the open-ended task of summarizing obligations to be reduced to a well-defined and circumscribed linguistic analysis task. ADEPT comprises a linearizer, which converts deeply nested sentences into a form that can be handled by standard parsers, a deontic sentence classifier trained on an annotated corpus of sentences drawn from representative policy documents, a semantic role analyzer, and other analytic tools for extracting and analyzing the deontic content of policy documents.

1 INTRODUCTION

Modern administrative states are regulated by statutes, regulations, and other authoritative legal sources that are expressed in complex, interconnected texts. Compliance with these rules is challenging for agencies, citizens, rule-drafters, and attorneys alike. For agencies, compliance requires understanding changes in federal laws, executive orders, and authoritative directives, policies, regulations, and standards. Simply identifying and summarizing these changes, which often originate from a multitude of sources, can be a burdensome drain on staff resources. The diversity of authoritative sources imposing requirements of a given nature is illustrated by the proliferation of cybersecurity requirements on U.S. federal agencies. Directives can be expressed in Executive Orders, Office of Management and Budget (OMB) circulars and memoranda, Department of Homeland Security (DHS) Binding Operational Directives (BODs), National Institute of Standards and Technology (NIST) Federal Information Processing Standards (FIPS), and Special Publications (SPs). Each agency must devote staff to monitor and review multiple streams of publications to identify changes affecting their cybersecurity profile (i.e., policies, practices, procedures, standards, and/or guidance). A similar monitoring task is required for all other areas within an agency where compliance is compulsory, such as privacy, health policy, and processing of sensitive information. An algorithmic process that automated the identification of sentences expressing obligations incumbent upon a given agency could significantly reduce the burden on staff having to review a large stream of documents. Such automated processes could provide agencies with early warnings of pending obligations, enabling them to better plan for implementation once the obligation is finalized.

A key observation of human performance on the document-monitoring task is that the summaries produced by staff typically focus on sentences that express *obligations*, i.e., that are characterized by *deontic modality*. This suggests that the tasks of monitoring and extracting directive sentences depend critically on the identification of such deontic sentences. We hypothesize that exploiting this insight will permit an important portion of the open-ended task of summarizing obligations to be reduced to a well-defined and circumscribed linguistic analysis task.

The remainder of this paper describes the design of a system for automated extraction of directives, ADEPT, and the evaluation of the critical deontic-sentence classification component. Section 2 presents examples of directives and explores the characteristics that distinguish directives from non-directives and different types of directives from one another. Section 3 discusses prior related work on modality classification, and the handling of nested directives, that is, sentences where dependent clauses or sentential complements share a common root clause is discussed in Section 4. Section 5 sets forth ADEPT's approach to identifying and classifying directive sentences, and Section 6 describes the use of semantic role labeling and frame instantiation to extract structured knowledge from sentences identified as directives. The implemented ADEPT architecture is described in Section 7, and Section 8 summarizes and outlines future efforts.

2 DIRECTIVE SENTENCES IN POLICY DOCUMENTS

ADEPT is based on an analysis of the work products of subject matter experts engaged in monitoring federal policy documents originating from the authoritative sources such as those listed in

In: Proceedings of the First Workshop on AI in the Administrative State, June 17, 2019, Montreal, QC, CA.

Section 1. Analysis of these sentences revealed that directives typically consist of expressions of obligations on the part of an agency or other government entity to perform or refrain from some specified actions, such as:

(1) Agencies must establish performance goals.

(2) Agencies are required to provide narrative responses regarding their risk management decision process.

(3) Each agency business owner is directed to ensure that 3DES and RC4 ciphers are disabled on mail servers.

(4) Chief Information Officers are to submit a report within 180 days.

These directive sentences can be viewed as illocutionary [3] or performative texts [22] that make a given action compulsory for a given government entity (i.e., the agency or a holder of a role within the agency). Frequently, as in sentence 1 above, directive sentences use modal verbs, such as "must", "shall", "may", and "should", as auxiliaries [21]. However, sentences 2–4 illustrate that obligations can be expressed without the use of modal verbs.

In addition to these *absolute*, i.e., *unqualified*, sentences, there are two other types of sentences that are important for some, but not all, applications.

First, some directives are *qualified* in the sense of expressing either *permission* or *weak necessity*, as in the following two sentences:

(5) Senior executives may consider delaying awarding new financial assistance obligations (permission).

(6) Agencies should establish and report other meaningful performance indicators and goals (weak necessity).

Second, some sentences merely report an obligation created by a different document, rather than creating an obligation themselves, such as:

(7) Section 1 of the Executive Order requires agency heads to ensure appropriate risk management.

We term such sentences *indirect obligation sentences*.

We exclude sentences from our set of directive sentences those that specify the details of an obligation created in a different sentence, e.g., by elaborating on the requirements of a work product obligation:

(8) Reports must enumerate performance goals.

We treat these sentences as non-directives because they provide details of obligatory actions but do not in themselves create an obligation for an agency or other government entity. We defer handling of these sentences to future applications.

In summary, we found that directive summaries extracted from policy documents by human experts typically have deontic force, which may be absolute, qualified, or indirect, depending on the construction of the sentence. We hypothesize that summaries consisting of these deontic sentences often closely match existing work products by agency personnel who currently monitor such documents and that summaries of this type may benefit agencies by enabling agency personnel to quickly identify the impact of new obligations, improving an agency's ability to comply in a complete and timely fashion.

3 RELATED WORK

Providing assistance to agencies in complying with complex regulatory and policy constraints is increasingly recognized as an important AI application. Typical examples include development of knowledge acquisition techniques to increase the agility in public administration [4] and information retrieval techniques optimized for regulatory texts [7]. This research has addressed both cross-document relationships among regulatory and statutory texts, such as network structure [15], and within-document analysis, such as discourse analysis of regulatory paragraphs [5] and parsing statutory and regulatory rule texts into a computer-interpretable form [24]. The work most closely related to the pragmatic objective of the current work is [18], which addressed sentential modality classification of sentences in financial regulation texts.

A number of previous research projects have addressed the general task of modal sense disambiguation (MSD) in legal and government texts. Marasović and Frank [16] developed a classifier for *epistemic*, *deontic*, and *dynamic* modal categories in English and German using a one-layer convolutional neural network (CNN) with feature maps and semantic feature detectors, reporting better results than with MaxEnt or a one-layer neural network. O'Neill et al. [18] combined a neural network with both legal-specific and more general distributional semantic model representations to distinguish among the deontic modalities *obligation*, *prohibition*, and *permission*. Wyners and Peters [20] used a rule-based approach to extract conditional and deontic rules from the U.S. Federal Code of Regulations. They found that this approach worked well for a specific set of regulatory texts, but its generality is unclear. Maat et al. [8] compared machine learning approaches to knowledge-based approaches for legal text classification in Dutch legislation, finding that while machine learning classifiers performed as well as the pattern-based model, the pattern-based approach generalized better than the machine learning model to new texts.

The modality classification task addressed by ADEPT differs from this prior work in that it focuses on the deontic distinctions relevant specifically for the task of extracting and summarizing the directives from administrative and policy documents, e.g., distinguishing deontic from non-deontic sentences, or identifying some subset of the categories of deontic sentences required for a particular application, such as just absolute and qualified obligations. As discussed below, ADEPT additionally addresses tasks both upstream from deontic sentence detection, such as linearization of nested directive sentences, and downstream, such as instantiation of obligation frames and conversion of instantiated frames into a structured form useful to agency personnel.

4 HANDLING NESTED DIRECTIVES

Authoritative administrative texts, including directives, regulations, and statutes, are often expressed in the form of nested enumerations, such as the directive set forth in Figure 1. Nested structures are characterized by multiple dependent clauses or sentential complements to common superordinate clauses. Such structures are intended to express complex rules and directives in a compact and comprehensible style by reducing textual redundancy. Human readers can easily understand the logical structure of such sentences

All agencies are **required** to:

1. Within 30 calendar days after issuance of this directive, develop and provide to DHS an "Agency Plan of Action for BOD 18-01" to:
   a. **Enhance email security by**:
      i. Within 90 days after issuance of this directive, configuring:
         - All internet-facing mail servers to offer STARTTLS, and
         - All second-level agency domains to have valid SPF/DMARC records, with at minimum a DMARC policy of "p=none" and at least one address defined as a recipient of aggregate and/or failure reports.
      ii. Within 120 days after issuance of this directive, ensuring:

Figure 1: A typical nested directive sentence. By itself, punctuation is insufficient to disambiguate whether the phrase in the box is a child of "Enhance email security …" or "Within 30 calendar days …". Either indentations or enumeration/itemization marks are required to resolve this ambiguity.

because the relationships among clauses are signaled by hierarchical relations between varying levels of enumeration symbols, punctuation marks, and varying indentation depths.

Unfortunately, parsers trained on standard treebanks, which are generally based on articles from news sources such as the Wall Street Journal, are unable to adequately process sentences with nested enumerations [17]. Thus, until domain-specific treebanks have been developed for legal texts which include nested sentences, it will remain necessary to convert such sentences into a format that is more amenable to conventional parsers.

One approach to simplifying the syntactic structure of nested enumerations is to convert them into a series of unnested sentences "by starting from the root of the tree and by concatenating, for each possible path, the phrases found until the leaves are reached" [10]. Each depth-first traversal of this tree is a simple (non-compound) sentence. We refer to this process as *linearization*. For example, the first sentence in a linearization of the nested sentence shown in Figure 1 is:

(9) All agencies are required to within 30 calendar days after issuance of this directive, develop and provide to DHS an agency Plan of Action for BOD 18-01 to enhance email security by within 90 days after issuance of this directive configuring all internet-facing mail servers to offer STRTTLS.

Linearization of regulatory and statutory text can be complicated by ambiguity in the scope of logical connectives that can arise from inconsistencies in expressing conjunction and disjunction in legal texts [1]. Nested directives, on the other hand, appear to generally be implicitly conjunctive, so linearization into a set of separate individual directives, each corresponding to a path in the depth-first traversal of the tree representing the logical form of the sentence, is generally consistent with the intended semantics of the original nested form.

As a practical matter, the greatest challenge in documents published in PDF (the primary format used by the agencies that we support) is determining the nesting level of each constituent clause with respect to surrounding clauses. Text extracted using standard tools, such as Apache Tika [2] and Tesseract [23], does not reliably retain the indentation depths of the original PDF. Punctuation marks often signal the nesting level, e.g., a clause that ends with a colon is to be followed by one or more subordinate (more deeply nested) clauses, and a period usually indicates a leaf node. However, there is an inherent ambiguity in sentences that follow a leaf node, such as the sentence in the box in Figure 1: "Within 120 days after issuance of this directive, ensuring:". Without either an unambiguous indication of indentation depth relative to surrounding clauses or an enumeration mark signaling a clear relationship to other lines of enumerated text, it is impossible to determine whether this sentence is (1) at the level of the sentence that starts "Within 90 days", (2) at the level of the sentence that starts "Enhance email security by:", or (3) the start of a new nested expression.

The lack of accurate indentation depths in text extracted from PDF documents and the ambiguity of the typical punctuation conventions suggest that the enumeration and bullet symbols and punctuation must be the source of nesting information. After all, these are generally unambiguous for human readers. Unfortunately, there is no canonical hierarchical practice of enumerations and bullets; document conventions vary not just among agencies but often within the same issuing agency as well from one document to the next. Enumeration and bulleting formats are sometimes applied inconsistently even within the same document. Our strategy is therefore to make an initial traversing pass through each document, recording the order of occurrence of each of a standard set of possible enumeration styles and conventions to establish a given document's hierarchical structure in each section. Each nested expression is then replaced with its linearized equivalent as determined from the hierarchy determined in the initial pass. The Appendix sets forth this procedure in more detail.

Our approach differs from Dragoni et al. [10], which mapped enumerated propositions onto a legal ontology to define the domain of directives and their constituent subparts, in using a concept-agnostic approach that may be better suited for domains in which directives are frequently revised, rescinded, or recontextualized in ways that may not be amenable to previous ontologies.

The extraction tools described below are intended to remove reference footnotes, HTML links, page numbers, and other extraneous information from within the span of single extracted sentences, but remaining bits of extraneous text create challenges for NLP processes downstream in our pipeline, such as POS and dependency parsing, event extraction, and modality detection. The last step of the linearization component therefore attempts to push these remaining items to the bottom of the linearized document as standardized endnotes.

5 DIRECTIVE SENTENCE CLASSIFICATION

Our working hypothesis is that policy-document summaries consisting of some or all categories of directive sentences described above can be a proxy for, assist in the creation of, or supplement manually-created compliance summaries. Thus, we focus on classifying sentences with respect to these directive sentence categories.

Table 1: The proportion of sentences of each of the 3 directive types and of non-directive sentences having a modal auxiliary.

Type	Ratio	Percent
Absolute	872/1626	54%
Qualified	552/875	63%
Indirect	60/158	38%
Non-directive	356/634	56%
Total	1840/3293	56%

5.1 Directive-Sentence Corpus

Unfortunately, none of the models or corpora developed in the prior work on sentence modality classification described above are directly applicable to our task. We therefore found it necessary to develop a new annotated directive sentence corpus based on U.S. executive-branch policy directives. Our initial focus was on OMB Memoranda and DHS Binding Operational Directives, for which we had examples of agency work products. We downloaded 5 years of OMB directives from the White House website.[1]

Each of the documents in the corpus was originally published in PDF format, usually with the first page scanned and signed. Each document was converted to plain text using the Apache Tika software package [2]. In parallel, each document was processed with Grobid [13] to identify elements such as headers and footers that can interrupt text that spans from one page to the next. The elements identified using Grobid were disinterleaved from the main text and concatenated at the end of each document.[2]

As described in Section 2, policy documents often contain complex sentences, including bullet-pointed lists and enumerations, which establish multiple distinct obligations. Accordingly, each nested sentence in the corpus was converted into a set of simple sentences using the linearization process described in Section 4. Each of the resulting sentences was then annotated according to the categories set forth in Section 2 by several annotators, including a subject-matter expert and a team of linguists.

The resulting set of 3,293 labeled sentences served as ground truth in the construction of the machine learning-based models described below. Roughly 95% of these sentences contain 10 tokens or more. Table 1 shows the proportion of sentences of each of the 3 directive types that have a modal auxiliary.[3] These ratios illustrate that the presence of modal auxiliaries is neither necessary nor sufficient for directives in this domain.

5.2 Evaluation of Deontic Sentence Classification

We applied the Weka [14] implementation of bagged random forests to the task of distinguishing among the four deontic categories of relevance to our task. Each sentence is converted to a feature vector consisting of n-grams and features derived from a dependency parse. As shown in Table 2, in an 80/20 train-test split of our corpus we achieved an F-score of 0.918 in distinguishing all directives from non-directives and a mean F-score of 0.816 across all four categories. A boosted decision tree model [11] using XGBoost [6] produced similar results.

P	R	F1	ROC Area	Class
0.762	0.954	0.847	0.949	Absolute
0.869	0.570	0.689	0.927	Qualified
1.000	0.346	0.514	0.862	Indirect
0.889	0.949	**0.918**	0.980	Non-Directive
0.843	0.829	**0.816**	0.950	Weighted Avg.

Table 2: Four-category deontic sentence prediction accuracy.

This experiment indicates that the deontic categories of relevance to our task can be distinguished by a model trained on a corpus of modest size. We anticipate increasing accuracy as we expand our annotated data set and as we refine the text extraction and linearization processes that provide input into the classifier.

6 SEMANTIC ROLE LABELING AND TEMPLATE INSTANTIATION

For many agency applications, the most useful representation of directives is often in the form of structured tables or spreadsheets summarizing multiple sentences. Analysis of representative work products indicated that the information of interest from each sentence includes the following:

- Actor - the agency or office to which the obligation applies
- Activity - the activity that is required of the Actor
- Object - the work product to be produced by the Activity, if any
- Time - any time-related qualification of the directed activity
- Manner - any non-time-related qualification of the directed activity
- Modal - the particular modal or other verb used to convey the deontic character of the expression, i.e., "must" vs. "may."

For each absolute or qualified directives, we instantiate a frame containing argument slots for each of the types of information above. For example, the instantiated frame shown in Table 3 summarizes the key information from the following directive sentence:

(10) Within 60 days of this Memorandum's publication agencies must update their list of non-governmental URLs.

Table 3: An instantiated directive template.

Actor	agencies
Activity	update
Object	list of non-governmental URLs
Time	within 60 days
Modal	must

The slots in the directive frame are a domain-specific adaptation of standard semantic roles. We use the Semantic Role Labeling model of AllenNLP [12] to assign Propbank semantic role labels [19] to directive sentences. We then use a set of simple heuristic rules for mapping these SRLs to the slots of our frames, e.g., a Propbank "ARG0" is generally the Actor, "ARG1" is generally the

[1] https://www.whitehouse.gov/omb/information-for-agencies/memoranda/
[2] Footnote texts must be retained because they sometimes contain directives.
[3] We plan to make this annotated corpus available to researchers in 2019 at http://mat-annotation.sourceforge.net/.

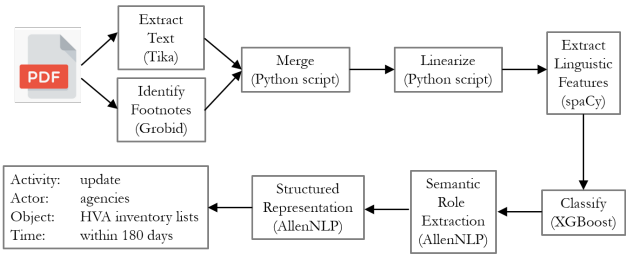

Figure 2: The directive sentence processing pipeline.

Object, and "Temporal" corresponds to the Time slot. Directives expressed without a modal verb ("All agencies are required to ...") will have no entry in the "Modal" field.

7 SYSTEM ARCHITECTURE

As illustrated in Figure 2, ADEPT's directive extraction and analysis tasks require a series of processing steps. We have adopted a modular architecture that can accommodate a variety of alternative components.

The first stage of the pipeline consists of concurrent calls to the APIs of the Tika and Grobid services offered by their respective Docker [9] containers. Tika outputs the PDF extraction as plain text whereas Grobid outputs the footnotes embedded in XML. The **merge** stage integrates this content and outputs a text file consisting of disinterleaved page content followed by all footnotes. The **linearizer** takes this text as input and outputs a text file containing one **linearized** sentence per line. The **linguistic feature extractor** converts each sentence into a feature vector of n-grams and features derived from a dependency parse.

An API call to the Docker container of the AllenNLP service is then made with a JSON file containing all sentences identified as being of the target deontic type or types (e.g., *absolute*). The AllenNLP output is passed to the template instantiation stage. The final output consists of CSV and HTML files that can be loaded into a spreadsheet or viewed through a web browser.

8 DISCUSSION AND FUTURE WORK

ADEPT illustrates how a document analysis task that imposes a significant burden to a wide range of agencies—directive extraction—can be addressed by deontic sentence classification in combination with nested sentence disambiguation and semantic role labeling. We anticipate that an ADEPT directive-extraction pilot will take place in mid-2019 with a representative U.S. federal agency.

Future work will relax ADEPT's current simplifying assumption that the directive content of policy documents can be determined by analyzing individual sentences divorced from their surrounding context. For within-document contextual information, we plan to introduce entity resolution and link connecting sentences that elaborate on an obligation with the obligation sentence to which they apply. To improve cross-document contextual information, we plan to develop techniques to detect and classify references to other documents, particularly statements that the current document rescinds directives from other policy documents.

Automated analysis of policy documents presents a rich set of text-analytic tasks but promises very significant rewards to both agencies and citizens. ADEPT represents an initial realization of this approach to improving the administrative state through modern computational linguistics techniques.

ACKNOWLEDGMENTS

The MITRE Corporation is a not-for-profit company, chartered in the public interest, that operates multiple federally funded research and development centers. This document is approved for Public Release; Distribution Unlimited. Case Number 18-4602.

REFERENCES

[1] L. Allen and C. Saxon. More IA needed in AI: Interpetation assistance for coping with the problem of multiple structural interpetations. In *Proceedings of the Third International Conference on Artificial Intelligence and Law*, pages 53–61, Oxford, England, June 25–28 1991.
[2] Apache tika - a content analysis toolkit. https://tika.apache.org/. Accessed: 2018-11-16.
[3] J. Austin. *How to do things with words*. Oxford U. Press, New York, 1962.
[4] A. Boer and T. van Engers. An agent-based legal knowledge acquisition methodology for agile public administration. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law*, ICAIL '11, pages 171–180, New York, NY, USA, 2011. ACM.
[5] A. Buabuchachart, K. Metcalf, N. Charness, and L. Morgenstern. Classification of regulatory paragraphs by discourse structure, reference structure, and regulation type. In *Proceedings of the 26th International Conference on Legal Knowledge-Based Systems JURIX*, University of Bologna, Bologna, Italy, November 2013.
[6] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.
[7] D. Collarana, T. Heuss, J. Lehmann, I. Lytra, G. Maheshwari, R. Nedelchev, T. Schmidt, and P. Trivedi. A question answering system on regulatory documents. In *Proceedings of the 31st international conference on Legal Knowledge and Information Systems (JURIX)*, 2018.
[8] E. de Maat, K. Krabben, and R. Winkels. Machine learning versus knowledge based classification of legal texts. In *Proceedings of the 2010 Conference on Legal Knowledge and Information Systems: JURIX 2010: The Twenty-Third Annual Conference*, pages 87–96, Amsterdam, The Netherlands, The Netherlands, 2010. IOS Press.
[9] DOCKER. https://www.docker.com/. Accessed: 2019-01-24.
[10] M. Dragoni, S. Villata, W. Rizzi, and G. Governatori. Combining NLP Approaches for Rule Extraction from Legal Documents. In *1st Workshop on MIning and REasoning with Legal texts (MIREL 2016)*, Sophia Antipolis, France, Dec. 2016.
[11] J. H. Friedman. Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 38(4):367–378, Feb. 2002.
[12] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. E. Peters, M. Schmitz, and L. Zettlemoyer. Allennlp: A deep semantic natural language processing platform. *CoRR*, abs/1803.07640, 2018.
[13] Grobid (or grobid) means GeneRation of BIbliographic data. https://grobid.readthedocs.io/en/latest/. Accessed: 2018-12-18.
[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.
[15] M. Koniaris, I. Anagnostopoulos, and Y. Vassiliou. Network analysis in the legal domain: a complex model for european union legal sources. *Journal of Complex Networks*, 6(2):243–268, 2018.
[16] A. Marasović and A. Frank. Multilingual modal sense classification using a convolutional neural network. In P. Blunsom, K. Cho, S. B. Cohen, E. Grefenstette, K. M. Hermann, L. Rimell, J. Weston, and S. W. Yih, editors, *Proceedings of the 1st Workshop on Representation Learning for NLP, Rep4NLP@ACL 2016, Berlin, Germany, August 11, 2016*, pages 111–120. Association for Computational Linguistics, 2016.
[17] L. Morgenstern. Toward automated international law compliance monitoring (tailcm). Technical report, LEIDOS, INC, 2014. AFRL-RI-RS-TR-2014-206.
[18] J. O'Neill, P. Buitelaar, C. Robin, and L. O'Brien. 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 2017, London, United Kingdom, June 12-16, 2017*, pages 159–168, 2017.
[19] M. Palmer, D. Gildea, and P. Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106, Mar. 2005.

[20] W. Peters and A. Z. Wyner. Legal text interpretation: Identifying hohfeldian relations from text. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA), 2016.
[21] The Plain Writing Act of 2010, 2010. 111th Congress H.R. 946.
[22] J. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, 1969.
[23] Tesseract ocr. https://opensource.google.com/projects/tesseract. Accessed: 2018-11-16.
[24] A. Wyner and W. Peters. On rule extraction from regulations. *Frontiers in Artificial Intelligence and Applications*, (235), January 2011.

9 APPENDIX: LINEARIZATION OF NESTED DIRECTIVES

FOR each document ingested by the linearizer:

> *Preprocess: Remove footnotes to prevent splitting of enumerated list elements or main body sentences during downstream processing later in the classification pipeline*

EXTRACT strings matching footnote format
STORE matching strings in References array
DELETE matching strings in their original positions
DELETE all multiple (n-1) vertical and horizontal spacing

> *Detect Document Section Boundaries: Identify positions of each document section to prevent enumerated elements from spanning multiple distinct lists.*

MATCH list of known section headers
STORE matches in partition along with starting offset position for each section in index
READ any enumerated lists in between section boundaries

> *Parse and Concatenate Enumerations: Map document hierarchical enumeration conventions against different symbol sets. Concatenate all directly subordinated sentence fragments with their subordinating fragments to form full (flat) sentences from the enumerated elements for downstream processing later in the classification pipeline.*

MATCH lines in each enumerated list within each section against enumeration symbol style list delimited by punctuation cues (Uppercase Roman Numerals, Lowercase Roman Numerals, Uppercase Letters, Lowercase Letters, Number Digits, Solid Bullet Points, Hollow Bullet Points)
STORE the sequential order (i.e., layers) of enumeration styles encountered to set document convention, where each layer begins with its own closet set of enumeration symbols
FOR lower-order layers
CONCATENATE lines recursively with all parent layers
TERMINATE upon reaching new paragraph with no enumeration symbol at the start of the line
ITERATE over all sections
WRITE to [FILENAME]_paths.txt file

> *Standardize Global Enumeration: Rewrite enumeration conventions to standard format (e.g. I.iii.B.a. → 1.3.2.1.)*

FOR all enumerated lists,
REWRITE each line's enumeration symbol with its corresponding digit based on the layer order and within-layer order
WRITE to [FILENAME]_trees.txt file

> *Post-Process Footnotes: Add previously extracted footnotes to the bottom of document*

APPEND footnote elements to bottom of the [FILENAME]_paths.txt file under the new section header "Footnotes"

Explicit interpretation of the Dutch Aliens Act

Specifications for Decision Support Systems and Administrative Practice

Robert van Doesburg†
Leibniz Institute
University of Amsterdam / TNO
Amsterdam, the Netherlands
robertvandoesburg@uva.nl

Tom van Engers
Leibniz Institute
University of Amsterdam / TNO
Amsterdam, the Netherlands
vanengers@uva.nl

ABSTRACT

This is a report of the explicit interpretation of the Dutch Aliens Act using the Calculemus method and the FLINT language. The method has been used before to make normative interpretations of regulations that form the basis for specific services to be delivered by governmental agencies and of legal cases. In this paper, the authors make an interpretation of an entire act, the Alien Act. We give an overview of methodical choices that enable the analyses of extensive sources of norms, and report on the results of the analysis.

CCS CONCEPTS

• Computing methodologies → Artificial intelligence → Knowledge representation and reasoning → Reasoning about belief and knowledge • Information systems → Information systems applications → Decision support systems → Expert systems • Applied computing → Law, social and behavioral sciences → Law

KEYWORDS

Knowledge acquisition, Legal engineering, Normative relations, Norm interpretation

1 Introduction

The Dutch Immigration Service (IND) was amongst the early adaptors of Artificial Intelligence. Using rule-based systems as of the early 1990s, they found that their, once advanced solution for deciding on alien cases, technically outdated in the early 2000s. The processes supported were grown too complicated and consequently too elaborative to maintain. This led to the awareness that the basis for the IND's processes and AI-supported decision-making and case-handling processed needed to be based on entirely new principles, i.e. Rule Governance. According to the Rule Governance principles systems like the one supporting the IND in its complex tasks, should be based on an aspect-oriented architecture (AOA) with a clear separation between the legal rules and the business requirements for supporting case-handling. In 2012 IND finished the implementation of the new rule-based information system INDiGO. The challenges the IND faced when it was designing and building INDiGO were:

1. Reduce the complexity of processes and systems.
2. Build a system that is flexible and agile in response to changes in sources of norms.
3. Build a system that is supporting professional employees and is not perceived as a straitjacket.

The development of INDiGO was a case study for the NWO-sponsored AGILE project. AGILE is an acronym for Advanced Governance of Information services through Legal Engineering. This resulted in several publication [7][13]. More recently INDiGO was used as a study case for a legal thesis on agile law making [14] (in Dutch).

Though INDiGO is doing fine as the information system for the IND, the ambition to build a system that is flexible and agile in response to changes in sources of norms is not fully achieved. At the same time, the need for flexibility and agility in relation to changes in sources of norms, traceability of the origins of norms is the system, and accountability on the compliancy of the system have become more important. The Dutch national government aims to have all public services to be available digitally. Furthermore, there is a debate going on requiring comprehensible explanation for all (automatic) decisions. In co-operation with the Dutch Tax and Custom Administration (DCTA) the Leibniz Institute and the PNA group, the IND has searched for adequate solutions [5][14].

This paper is about the legal engineering aspects of INDiGO. We show an example of the explicit interpretation of the Dutch Aliens Act that aims to:

1. Make explicit interpretations of sources of norms that can be used in multidisciplinary teams consisting of lawyers, policy advisors, administrative workers and knowledge engineers.
2. Make comprehensive, high-level interpretations of sizable amounts of sources of norms.
3. Enable a modular approach that allows adding detailed interpretation at a later time or adding links to interpretation of other sources of norms.
4. Enable structured debate on disagreements on interpretations of sources of norms and the application thereof in specific cases.

In: Proceedings of the First Workshop on AI in the Administrative State, June 17, 2019, Montreal, QC, CA. B)2019

In this paper, we will discuss the first three aims. For more on structured debate on disagreements, see [3].

In Section 2, we will give a short overview of the results of early work in our quest for explicit interpretations of sources of norms and its theoretical basis. In Section 3, we will give a short introduction of the Calculemus method and the FLINT language for the explicit interpretation of sources of norms. In Section 4, we will present the results of the analysis of the generative norms in the Dutch Aliens Act. In Section 5, we will discuss results and future work.

2 Early work and Theoretical Framework

In the development of INDiGO the acquiring and modelling of legal knowledge was a bottleneck. Legal experts and knowledge engineers had to work together to make sound models. However, while the legal experts could not see how a piece of text was not sufficient to be machine interpreted, the semantic engineers could not understand the interpretations of the legal experts. Knowledge engineers proposed a method that interprets norms as descriptions of behavior that is either allowed or forbidden [13]. This is essentially a deontic perspective on norms.

The approach succeeded in so far, that using this conceptualization the rule-base of INDiGO was created, that functions reasonably well. The goal of creating a single-source-of-knowledge on norms and rules, however, was not achieved. In our opinion this has two main causes:

1. The pursuit of completeness of the legal framework.
2. The choice for a deontic approach.

By striving to analyze 500+ laws and regulations in order to create a complete normative knowledge base, an impossible task was created. A modular approach that starts with a high-level interpretation of the core of the sources that regulate the work of the IND, would have been better [14]. In this paper we present such an approach.

The choice for a deontic approach neglects the importance of the regulation of the power to act. Power is a under specified concept in AI and Law [16].

In 1931 Kocourek [12] stated that there are different opinions on the number of fundamental normative relations, but that there was nobody that believed that there are more than four. People who follow the deontic approach believe that the fundamental legal relation is the *claim-duty* relation. This means, that in their opinion all normative positions and relations can be expressed using deontic concepts, e.g. Herrestad [10]. Kocourek himself believed that there are two fundamental normative relations: the *claim-duty* relation and the *power-liability* relation. According to Kocourek, Hohfeld is the most important proponent of four fundamental normative relations: *claim-duty* relation, *power-liability* relation, *liberty-no claim* relation, and *immunity-disability* relations [11]. In practice, there is little difference between the position that there are two or four fundamental normative relations. There is not much difference to claiming a *liberty-noclaim* relation is fundamental, or that it is the same as an absent *claim-duty* relation. Therefore, we work with a model that is based on two fundamental normative relations.

In Section 3 we will present an approach that interprets sources of norms from an action perspective. We will do so by introducing the Calculemus method for building explicit representations of normative relations, and the FLINT language to express these representations.

For more on the theoretical base of FLINT, see [4].

3 The Calculemus method

In the last years we have been developing a method that aims to make interpretations of sources of norms that can be used:

1. to make specifications for decisions being taken by machines and people in administrative organizations,
2. to support the grounding of decisions made in courts, and
3. to support the implementation and evaluation of policies in large organizations.

This work has resulted in a method to address questions related to norms between people and in organizations [1][2][3][4]. The goal of this method is to create a method for the interpretation of written or spoken sources of norms in natural language, resulting in specifications for normative multiagent systems that can be used by humans and machines.

For the expression interpretation of sources of norms, we are developing the Formal Language for the INTerpretation of sources of norms (FLINT). This is a now semi-formal language that is evolved from working on real-life cases. The language consists of three frames. One to describe normative actions performed by an actor that results in normative changes addressed to an agent that is either receiving the results or is an interested party. This frame is called an *act type frame*, because it describes all the aspects of the function that changes a state due to a normative action performed by an agent. We make a distinction between the action of the agent and all that is necessary to achieve an effect, i.e. a transition to another normative position. The *act type frame* is presented in Table 1.

The *act* consists of the *action* of *actor*, an *object* that is acted upon and is submitted to a transition because of that *action*. The *act* is *valid* if it is performed in a state that meets a *precondition*. If the *action* is performed and the *precondition* is met, the action will result in *normative facts*, and/or *normative relations* being either *created* or *terminated*. The *act type frame* describes a *power-liability* relation in which the *actor* holds a *power* to perform a certain *action* with a functional effect that comes to existence if the *action* is *valid*, i.e. if the *precondition* is met. The *interested party* holds a *liability* related to the action of the actor.

The *act frame* is a concept that exist only in institutional reality. The *act frame* and all its components are constituted by giving additional meaning to facts in social reality by qualifying them as normative components. As a result, *institutional reality*

does not have a procedural perspective. Time is only relevant in institutional reality because of the time interval in which sources of norms are valid, the time interval in which facts occur in social reality and the time that facts in social reality are qualified as normative, or institutional facts.

Act frame	<<name of the *act frame*>>
Action	Action that causes the transition of an object
Actor	Agent role that is allowed to perform action
Object	The object acted upon
Recipient / Interested Party	Agent role having a normative relation with the actor concerning his action
Precondition	Set of conditions that must be met to allow the action of the actor
Creating postcondition	Facts or normative relations created by action of the actor
Terminating postcondition	Facts or normative relations terminated by action of the actor
References to sources	Reference to the source of the act type, including information on version

Table 1: The Institutional Act Type Frame and its Constituents

The results of *acts frames* can be *facts* that are *created* because of the transition of the *object*. The *object* itself can be seen as a *fact* in the role of the *object* of *action*. For example: if an application for a permit is positively decided upon, it results in the *creation* of a permit, and the *termination* of the application, because it is not desirable to take a new decision on the application before withdrawing the first one. A *normative act*, represented by an *act frame*, can also create two types of normative relations: potestative relations and deontic relations. Potestative relations can be used to create new *act frames*, e.g. the decision that an official gets the *power* to execute *normative actions*.

Duty frame	<expression of the *duty* (future act)>
Duty holder	Agent role holding the *duty*
Claimant	Agent role holding the *claim*
Creating institutional act	The *normative act(s)* that creates the *claim-duty* relation
Enforcing institutional act	The *normative act(s)*that the *claimant* can use to enforce the satisfaction of the duty in case the *duty holder* renounces a duty
Terminating institutional act	The *normative act(s)* that satisfies the claim-duty relation (effectively terminating it)
References to sources	References to fragments of sources of norms for all frame elements, including information on version

Table 2: The Duty Frame and its Constituents

The deontic normative relation consists of a *duty*, or obligation, that is in effect the state in which an institutional act that ought to be performed in the future, or ought to have been performed in the past in case of a violation of the duty. A *duty frame* consists of an expression of the *duty* involved. It has a *duty holder* and a *claimant*. The *duty frame* also has one or more *creating act frame(s)* that can create the *duty*, *enforcing act frame(s)* that can be used to enforce the satisfaction of the *duty* in case the *duty holder* renounces his duty, and *terminating* (or satisfying) *act frame(s)* that effectively terminates the *claim-duty* relation the *duty frame* is an expression of.

In case a *duty holder* is of the opinion that he does not have a *duty*, he can *claim* a privilege or *liberty* towards the *claimant*, using an appropriate *act frame*. The claimant and the *duty holder* now have a conflict on the question whether a *claim-duty* relation exists, or not. The argument usually will be about the question whether the *normative act* that *created* the *duty* was *valid*, or whether or not the *normative act* that was supposed to *terminate* the *duty* was *valid*. Since this question is about the presence or absence of a *claim-duty* relation there is no need for a separate frame for *liberty-no claim* relations. The *duty frame* is presented in Table 2.

Fact frame	[expression of fact]
Function	Boolean function expressing the condition that makes a fact true, or an arithmetic function, e.g. for calculating amounts of money
References to sources	Reference to the source of the fact type, including information on version

Table 3: The Institutional Fact Frame and its Constituents

The third frame of the FLINT language concerns the *fact type frame*. The *fact frame* that can be used to make detailed statements on the *precondition* of an *institutional act*. The precondition consists of a function of institutional facts connected by Boolean or arithmetic operators. Every *institutional fact* in the function of an *institutional fact frame*, can be the subject of a new *fact frame*. The level of detail that is pursued depends on the purpose of the analysis. The *institutional fact frame* is presented in Table 3.

4 Analysis of the Aliens Act

In 2018 a FLINT analysis was made for students and highly skilled workers that want to reside in the Netherlands, using the Calculemus method. Relevant sources of norms were collected, and explicit interpretations were made using the FLINT language. The experiment showed that it was possible to make a modular analysis of specific tasks of the IND using a middle-out approach, thus solving the gridlock caused by striving to completeness that was one of the causes for the unsuccessful attempt to create a single point of truth on normative knowledge for the IND during the development of INDiGO.

The results of this experiment also showed that the core concepts in the INDiGO information system, e.g. qualifications, criteria and evidence, did not, or no longer, reflect the specifications of sources of norms. Since the start of the INDiGO program, the original model, based on qualifications (rights that people want to qualify for), criteria (that should be met in order to be fit to qualify for a qualification), and evidence (with which

one can sufficiently proof that a criterium is met) were contaminated because practical solutions were chosen to solve urgent requests of users of the systems. The architectural principles of the original system, as described in [7] were not properly guarded.

The analysis of the sources of norms for specific products, lead to the conclusion that there was a need for administrative specifications that combine a normative perspective with the practical perspective of administrators of the Alien Act. The question at hand, was whether it is possible to use the Calculemus method and the FLINT language to analyze larger amounts of sources of norms, without getting trapped by the quest of completeness. In legal analysis there is always the risk to get lost in exotic details, or possible exceptions: very interesting from a legal perspective, but irrelevant from an administrative perspective. We tried to bypass this trap by limiting the analysis to acts, i.e. *act frames*, that are grounded in the Aliens Act. The ratio of this choice is that by looking for *act frames*, we would be able to address all causes for the generation of new facts and new normative relations. The generation of normative facts or relation always requires an act.

The Dutch Aliens Act is one of the two main acts that are administrated by the IND.[1] The Aliens Act regulates more then 80% of the work done by the IND. The question we want to answer using the Calculemus method is: "What Act Frames Can Be Retrieved from the Aliens Act?"

In order to answer that question, we have taken 5 steps:

1. Choose a way to split up the Aliens Act into separate containers of knowledge.
2. Classify constituents of the Aliens Act that contain *act frames*.
3. Make classes of *act frames* that are found in the Aliens Act.
4. Assign *act frames* to the IND.
5. Make guidelines for elaborating pre- and postconditions.

4.1 How to Split Up the Aliens Act in Separate Containers of Knowledge?

In April 2001 the first version of the Aliens Act came into effect. Since than numerous changes have been made to the original text. Every change was separately published by the Dutch government. The versions of the Aliens Act on the web [8] (in Dutch) is an aggregation of the original text and the official publication of all changes to the Aliens Act. For this paper the version that was valid starting July 28, 2018 was used. The full text of the law was divided into clauses, subordinate clauses, and components of enumerations. The English translation of the Dutch legislation in this paper are unofficial translations by the authors. The English designations used are based on the guidelines of the European Union [6]. The reason for decomposing the text of the law, is to make a set of components that are minimal meaningful units.

[1] The IND is responsible for immigration (regulated by the Aliens Act), and naturalization (regulated by the Dutch Nationality Act).

In Table 4 we show an example is for the decomposition of Article 14 (1) Aliens Act. The decomposition of the Aliens Act into clauses is laid down in a spreadsheet that contains more detailed information of every component of the source, including separate versions of the same component valid at different periods of time. This decomposition will be used as a specification for future tooling for the decomposition of sources or norms. Emile de Maat [15] built a prototype for automatically decomposing Dutch legislation and improving the possibilities for making identifiers for words, or groups of words, in sources of norms. Unfortunately, this prototype was never taken in production. For the purpose of this paper the decomposition of the Aliens Act was done by hand. De Maat's did experiment with several methods for the interpretation of norms, but did not come to a satisfying method for the interpretation of sources of norms in natural language.

Reference	Text	Valid since (yyyymmdd)
Art. 14 (1)	1. Our Minister is authorized:	20010401
Art. 14 (1)(a)	a. to grant, reject, or disregard the application to provide a temporary regular residence permit;	20010401
Art. 14 (1)(b)	b. to grant, reject, or disregard the application of the extension of the period of validity;	20010401
Art. 14 (1)(c)	c. to change a temporary regular residence permit, on application or ex officio, due to changed circumstances;	20130601
Art. 14 (1)(d)	d. to revoke a temporary regular residence permit;	20010401
Art. 14 (1)(e)	e. to grant, or to extend the period of validity of a temporary regular residence permit ex officio.	20130601

Table 4: Decomposition of Article 14 (1) Aliens Act

The Aliens Act consists of 1.387 components, of which 1.370 constitute the body of the law. The body of the law consists of 274 structural components, i.e. titles of chapters, divisions, sections, and titles of articles, leaving 1.096 components that contain sources of norms related to immigration policies.

4.2 Which clauses of the Aliens Act can be interpreted as act frames?

An act frame is a classification for a clause, or a combination of clauses, describing a normative act: an action, performed by an actor, on an object, while a precondition is met, with a result and an interested party. So, the question is, which clauses of the Alien Act can be interpreted as being part of an act frame?

Usually a sentence in a source of norm, e.g. the Aliens Act, that can be interpreted as an act frame, contains an action, an actor and the object that is acted upon. Though, not always in the same clause. In Table 4 you can see that the main clause of

Article 14 (1) contains an actor (Our Minister[2]), while the action, and object acted upon can be found in Article 14 (1)(a). That the interested party of this act frame is the alien that submitted the application, can be derived from Article 8 Aliens Act, in which it is stated that an alien has lawful residence in the Netherlands if he has a residence permit as mentioned in Article 14 Aliens Act.

The interpretation of the Aliens Act using *act frames* is done by selecting clauses that contain an action. A preliminary name for the act frame is connected to the clause. Then, the source text is examined in order to find the actor, the object and the interested party of the act frame at hand. The exact words from sources of norms are used to express components of the *act frame* and to adjust the name of the *act frame*, if necessary. The precondition and postcondition (results) of the *act frame* are left open for further interpretation at a later time. These parts of the act frame are more complex because precondition and postcondition may be a composition of multiple components from multiple sources.

There are three main groups of *act frames* we found in the Aliens Act. The first group concerns *act frames* that create additional rules, e.g. the creation of rules by order in council, e.g. in Article 2a (2) main clause and under (a) Aliens Act:

> By or pursuant to an order in council:
> a. further rules are laid down regarding natural persons and organizations that can act as sponsors;

The second group concerns acts that give authorities to the Minister of Justice and Security, or to officials that act in his name, e.g. Article 2c (2) main clause and under (a) Aliens Act:

> Our Minister is authorized:
> a. to grant, reject or disregard the application for recognition as sponsor,

The third group concerns *act frames* that give authorities to individuals that are an interested party in relation to the Aliens Act, e.g. Article 3 (4) Aliens Act, second sentence:

> A decision to refuse entry to the Netherlands, that has already been taken, will lapse as from the time at which the alien indicates that he wishes to submit an application as referred to in the third Paragraph.

Here an alien has the power to lapse a decision to refuse the entry to the Netherlands by applying for temporary asylum residence permit, i.e. the application mentioned in Art. 3 (3) Aliens Act.

Of the 1.096 clauses of the body of the Aliens Act 250 contained one or more *act frames*. A clause can be the source of multiple *act frames* when a clause contains multiple actions, like the clauses in Article 14 (1) Aliens Act, see Table 4, that represent *act frames* to grant, reject, or disregard the application to provide a temporary regular residence permit.

In the Aliens Act we found a maximum of four *act frames* in one clause. In total, we found a total 428 *act frames* in the Aliens Act.

[2] The fact that 'Our Minister' is the 'Minister of Justice and Security' is laid down in Article 1 Aliens Act.

4.3 Categories of Act Frames in the Aliens Act

The *act frames* in the Aliens Act can be divided in multiple categories. Table 5 gives an overview of the number of items for every constituting element.

Concepts	Number of Items
Actor	12
Action	96
Object	177
Interested Party	29

Table 5: Number of Items per Constituting Element of the Act Frames in the Aliens Act

In order to focus on the *act frames* for which the IND is responsible, we zoom in on the actors for whom a role is laid down in the Aliens Act, see Table 6. The actor charged with most of the acts described in the Alien Act, is the Minister of Justice and Security. These acts include the acts performed by the IND in name of the Minister of Justice and Security. The Dutch Government is involved in slightly less different acts. It concerns the authority to make additional rules by or pursuant to order in council on a large array of specific issues.

Actors	Number of Acts
Minister of Justice and Security	178
Government	161
Official Charged with Border Control	30
Official Charged with the Supervision of Aliens	25
District Court	13
Alien	10
Carrier	4
Administrative Authorities	2
Administrative Law Division of the Council of State	2
Chief of Police	1
Officer in Command of the Royal Netherlands Marechaussee	1
Sponsor (of the Alien)	1

Table 6: Actors in the Aliens Act and the Number of Acts Assigned to them

Officials charged with border control and supervision are mentioned separately because they not only take decisions in writing in name of the Minister of Justice and Security but are also authorized to use physical force for all kinds of acts, including imprisonment. The Commander of the Marechaussee is in charge of border control, the Chief of Police in charge of the supervision of aliens. The District Court is responsible for deciding on legal conflicts related to the Aliens Act. The administrative law division of the Council of State is responsible for the administration of appeals. Carriers responsible for illegally transporting people into the country can be forced to

transport people to a place outside the Netherlands free of charge. Aliens and sponsors can perform acts related to their procedures to gain rights regulated by the Aliens Act.

4.4 Which Act Frames are Assigned to the IND?

The Minister of Justice and Security is responsible for tasks he performs personally, but also for a wide array of tasks carried out in his name by officials. The tasks of the IND are amongst these.

Of the 178 tasks assigned to the Minister, 106 are assigned to the IND. These tasks can be divided into 5 main categories:

1. Residence permits and recognized sponsorship
2. Proof of lawful residence
3. Deadlines for taking decisions
4. Collecting and using biometrical data
5. Data and Knowledge Management

The result of this exercise is a complete set of actions the IND can perform based on the Aliens Act, divided in groups related to specific tasks and products. In Section 4.5 we show how the pre- and postconditions of *act frames* are constructed.

4.4.1 Residence permits and recognized sponsorship. The category residence permits and recognized sponsorship contains the core task of the IND. It contains 83 *act frames* concerning 45 objects. The IND administers the providing of four types of residence permits[3]:

1. Provisional residence permit (7 objects, 15 *act frames*)
2. Regular residence permits (12 objects, 25 *act frames*)
3. Residence permits based on European legislation (3 objects, 5 *act frames*)
4. Asylum residence permits (11 objects, 21 *act frames*)
5. Return visa (6 objects, 10 *act frames*).

The provisional residence permit is more like a visa than a residence permit. It is a visa to enter the Netherlands for people that apply for a regular residence permit. The regular residence permit is a residence permit that assigned for other reasons than for the purpose to give asylum. Regular residence permits are granted for specific purposes, e.g. family life, work or study.

In this paper, we will discuss the acts related to the granting, rejecting and disregarding of regular residence permits. The other categories are administered using technically comparable *act frame*, although there are considerable differences from a legal perspective. We will address these differences at another paper.

4.4.2 Proof of lawful residence. The IND is obliged to provide lawful residing aliens with a document or a written statement that proves lawful residence. For this task there are 5 objects and 5 *act frames* because for this category the only action is: to provide.

4.4.3 Deadlines for taking decisions. The Aliens Act requires the IND to decide on application within a limited period of time. There are separate time limits for every category of applications. Also, for all categories there are rules that regulate the possibility to extend the deadline for a decision. After the related *act frames* where extracted from the Aliens Act, we noticed the IND information architecture did not include decisions to extend the term available for deciding on application. Where the IND intends to use rule-based decisions for all tasks its primary process, decisions for extending deadlines were, until now, never seen as separate decisions. There are 10 different objects, 9 of which are related extending deadlines for taking decisions, and the other one related to making the decision known to interested parties.

4.4.4 Collecting and using biometrical data. The IND collects fingerprints and facial images of aliens for identification purposes. There are separate *act frames* for requesting biometrical data, collecting it, providing it to other agencies, receiving it from other agencies, and for comparing the fingerprints of aliens with fingerprints stored in a document of an administration. There are 6 objects and 6 *act frames* in this category.

4.4.5 Data and knowledge management. The Aliens Act describes how the IND ought to process personal data. The IND can designate administrations acquiring data concerning aliens and their legal procedures. There are *act frames* for requesting and acquiring data. There is an *act frame* for maintaining an aliens administration. And there are *act frames* for attaching written documents to applications. There are 6 objects and 6 *act frames* in this category.

4.5 Elaborating Pre- and Postconditions of Applications for Regular Residence Permits

Determining the pre- and postconditions of an *act frame* is not a straight-forward procedure. While the other elements of the *act frame* are singular concepts. The *act frame* can only concern one action and has one actor. However, the pre- and postcondition of an *act frame* may consist of complex statements. We will show some examples of complex preconditions and the use of *fact frames* for expressing the details of a complex precondition. The postcondition consists of one or more elements that are created or terminated if an act is valid.

We start with the representation of the *act frame* representing the rejection of an application to grant a temporary regular residence permit, see Table 7. We do so, because the Aliens Act contains specific grounds for rejecting an application to grant a temporary regular residence permit, or to disregard it, but not for granting it. The rules for granting an application must be derived from the inability to reject or disregard it.

Art. 66a (6) Aliens Act states that an alien that has a travel ban or has been signaled for the purpose of refusing entry,

[3] Permits for long-term residents of the European Union are, strictly speaking, a separate category, because they are based on European regulations. But for the purpose of this paper, they are included in the category regular residents permits.

cannot have a valid residence permit, asylum nor regular. The same goes for aliens that have been pronounced undesirable based on art. 67 (3) Aliens Act.

Act frame	<<rejecting a temporary regular residence permit>>
Action	[reject]
Actor	[Minister of Justice and Security]
Object	[application to provide a temporary regular residence permit]
Recipient / Interested Party	[alien]
Precondition	([alien has a travel ban or has been signaled for the purpose of refusing entry] OR [alien has pronouncement of undesirability] OR NOT [application contains purpose of residence] OR NOT [alien has a provisional temporary residence permit] OR NOT [alien has a valid border-crossing document] OR NOT [interested party has sufficient, independent, long-term means of support] OR [alien constitutes a threat to public order or national security] OR NOT [alien is willing to cooperate in a medical examination of a disease designated by the Public Health Act or to undergo medical treatment for such a disease] OR [alien has performed any work in violation of the Aliens Labor Act] OR NOT [alien meets the restriction related to the purpose of residence] OR NOT [alien has sufficient knowledge of the Dutch language and Dutch society] OR [alien has provided incorrect data or has withheld data] OR NOT [alien has only resided in the Netherlands on the basis of Article 8 Aliens Act] OR NOT [sponsor has submitted a statement for the purpose of the intended residence of the alien]) AND NOT [adverse consequences of a decision may not be disproportionate to the purposes to be served by the decision]
Creating postcondition	[decision to reject application to provide a temporary regular residence permit]
Terminating postcondition	[application to provide a temporary regular residence permit]
References to sources	Art. 14 (1) Aliens Act, main clause and under (a)

Table 7: Act frame for rejecting a temporary regular residence permit

The next eleven possible grounds for rejecting a residence permit can be found in art. 16 (1) Aliens Act. This article contains a full set grounds for the rejection of the application of temporary regular residence permits. These grounds are only relevant for temporary regular residence permits, not for permanent permits, or for asylum permits. Every of these grounds has several exceptions, e.g. art. 17 (1) Aliens Act contains a set of exceptions for the condition that an alien must have a provisional temporary residence permit, these exceptions, e.g. based on nationality, are not discussed in this paper.

The last condition, i.e. that a decision may not have adverse consequences that are disproportionate to the purposes to be served by the decision, is not to found in the Aliens Act, it is a general condition for decisions to be taken by administrative authorities laid down in the General Administrative Law Act (GALA) [9]. GALA provides a framework for Dutch administrative acts. In this paper, we will not go into the details of this matter. The fact that the result of the granting, rejecting or disregarding of an application for a temporary regular residence permit is a decision, is a result of the definition of the concept 'application' as the request by an interested party to take a decision (art. 1:3 (3) GALA).The application itself is terminated by deciding on it. If a decision has been taken on an application, it is not possible to make additional decisions on the same application, this is only possible after revoking the decision, as is regulated in GALA.

Disregarding of the application to provide a temporary regular residence permit, see Table 8, is a rather simple *act frame*, it can only be done on the ground that fees due for the settlement of to grant a temporary regular residence permit have not been paid (art. 24 (2) Aliens Act, third sentence). There is also a possibility to disregard application based on art. 4:5 GALA, because this is an act that is not based on the Aliens Act, it is not discussed in this paper.

Act frame	<<disregarding a temporary regular residence permit>>
Action	[disregard]
Actor	[Minister of Justice and Security]
Object	[application to provide a temporary regular residence permit]
Recipient / Interested Party	[alien]
Precondition	NOT [fees due for the settlement of to grant a temporary regular residence permit have been paid]
Creating postcondition	[decision to disregarding application to provide a temporary regular residence permit]
Terminating postcondition	[application to provide a temporary regular residence permit]
References to sources	Art. 14 (1) Aliens Act, main clause and under (a)

Table 8: Act frame for disregarding a temporary regular residence permit

This leaves the granting of a temporary regular residence permit, see Table 9. Art. 26 (1) Aliens Acts states that a regular residence permit is granted from the day on which the alien has demonstrated that he meets all conditions, but not earlier than from the day on which the application was received.

Art. 26 (1) is the only condition mentioned in the Aliens Act for granting a temporary regular residence permit. We will elaborate the *fact frame* [regular residence permit is granted from the day on which the alien has demonstrated that he meets all conditions] shortly, but first we will address the other elements in the precondition, and the postcondition of this *act frame*.

The other elements in the precondition —that an alien that is granted a temporary regular residence permit may not have a travel ban or have been signaled for the purpose of refusing entry, or have been pronounced undesirable based on art. 67 (3) Aliens Act— follow from the *act frame* on the rejection of the residence permit.

Act frame	<<granting a temporary regular residence permit>>
Action	[grant]
Actor	[Minister of Justice and Security]
Object	[application to provide a temporary regular residence permit]
Recipient / Interested Party	[alien]
Precondition	[regular residence permit is granted from the day on which the alien has demonstrated that he meets all conditions] AND NOT [residence permit granted earlier than from the day on which the application was received] AND NOT [alien has a travel ban or has been signaled for the purpose of refusing entry] AND NOT [alien has pronouncement of undesirability]
Creating postcondition	[decision to grant an application to provide a temporary regular residence permit]; <granting a temporary regular residence permit under restrictions>; <determine the period of validity of the regular residence permit>; <provide the alien with a document proving lawful residence>
Terminating postcondition	[application to provide a temporary regular residence permit]
References to sources	Art. 14 (1) Aliens Act, main clause and under (a)

Table 9: Act frame for granting a temporary regular residence permit

The postcondition of the granting of a permit has more elements than the rejecting or disregarding of it. Apart from the creation of decision to grant an application for the provision of a temporary regular residence permit and the termination of the application to provide one, several duties are created. The duties follow from art. 14 (3) Aliens Act, where it is laid down that a temporary regular residence permit is granted under restrictions related to the purpose of residence. Art. 14 (4) Aliens Act requires the determination of the period of validity of the residence permit, that may not exceed a period of 5 years, and Art. 9 (1) Aliens Act requires that a document proving lawful residence is provide to the alien that is granted a residence permit, i.e. to provide the document that is the residence permit. After elaborating the *fact frame* [regular residence permit is granted from the day on which the alien has demonstrated that he meets all conditions] we will show how these duties can be represented in FLINT.

Fact frame	[regular residence permit is granted from the day on which the alien has demonstrated that he meets all conditions]
Function	[alien has demonstrated that he meets all conditions of the regular residence permit] AND [day on which alien has demonstrated he meets all the conditions for a regular residence permit] AND NOT [day on which alien has demonstrated he meets all the conditions for a regular residence permit lies before the day the application was submitted]
References to sources	Article 26 (1) Aliens Act

Table 10: Fact frame for determining whether a 'regular residence permit is granted from the day on which the alien has demonstrated that he meets all conditions'

Fact frame	[alien has demonstrated that he meets all conditions of the regular residence permit]
Function	[alien allows himself to be photographed and to have his fingerprints taken] AND ([alien meets the conditions to provide a temporary regular residence permit provision] OR [alien meets the conditions for extending a temporary regular residence permit] OR [alien meets the conditions for changing a temporary regular residence permit] OR [alien meets the conditions to provide a permanent regular residence permit])
References to sources	Article 26 Paragraph 1 Aliens Act

Table 11: Fact frame for determining whether a 'alien has demonstrated that he meets all conditions of the regular residence permit'

The fact [regular residence permit is granted from the day on which the alien has demonstrated that he meets all conditions], see Table 10, is based on Art. 26 (1) Aliens Act.

Fact frame	[alien meets the conditions to provide a temporary regular residence permit]
Function	[application contains purpose of residence] AND [alien has a provisional temporary residence permit] AND [alien has a valid border-crossing document] AND [interested party has sufficient, independent, long-term means of support] AND NOT [alien constitutes a threat to public order or national security] AND [alien is willing to cooperate in a medical examination of a disease designated by the Public Health Act or to undergo medical treatment for such a disease] AND [alien has not performed any work in violation of the Aliens Labor Act] AND [alien meets the restriction related to the purpose of residence] AND [alien has sufficient knowledge of the Dutch language and Dutch society] AND NOT [alien has provided incorrect data or has withheld data] AND [alien has only resided in the Netherlands on the basis of Article 8 Aliens Act] AND [sponsor has submitted a statement for the purpose of the intended residence of the alien] AND [fees due for the settlement of to grant a temporary regular residence permit have been paid]
References to sources	Article 26 Paragraph 1 Aliens Act

Table 12: Fact frame for determining whether a 'alien meets the conditions for granting a temporary regular residence permit provision'

First we split the fact in three, resulting in the fact [alien has demonstrated that he meets all conditions of the regular residence permit], the fact [day on which alien has demonstrated he meets all the conditions for a regular residence permit], and [day on which alien has demonstrated he meets all the conditions for a regular residence permit lies before the day the application was submitted]. The fact [alien has demonstrated that he meets all conditions of the regular residence permit] is then elaborated in Table 11. The other two facts need an arithmetic function to determine whether they are true or false.

In Art. 106a (1) Aliens Act, first sentence, the condition that the alien assists with providing a photograph and fingerprints to being taken. Additionally, from Art. 14 (1) Aliens Act it follows that there are four act types related to regular residence permits, with separate pre- and postconditions. For determining the conditions for granting a temporary regular residence permit we need to zoom in on fact [alien meets the conditions for granting a temporary regular residence permit], see Table 12.

The Boolean function that makes the fact [alien meets the conditions to provide a temporary regular residence permit] true consists of the combination of the precondition of rejecting the application and disregarding it, see Table 12.

4.6 Satisfying Duties Created by Granting a Regular Residence Permit

The duty to grant a temporary regular residence permit under restrictions is created by granting the permit. Enforcing of the duty is possible by submitting a letter of objection. The duty can be fulfilled or terminated by a whole set of acts that are described in the Aliens Decree, an Order in Council that contains the majority of the rules that can be made based on *act frames* in the Aliens Act. For every restriction there is a separate *act frame*, with a specific precondition and a specific result. There are more than hundred separate restrictions, two of them are shown in Table 13.

Duty frame	<granting a temporary regular residence permit under restrictions>
Duty holder	[Minister of Justice and Security]
Claimant	[alien]
Creating institutional act	<<grant an application for a temporary regular residence permit>>
Enforcing institutional act	<<submit a letter of objection>>
Terminating institutional act	<<grant a temporary regular residence permit under restriction: Residence as a family member with (name). Work freely allowed. Work permit not required.>>; … << grant a temporary regular residence permit under restriction: Residence for non-temporary humanitarian grounds. Work freely allowed. Work permit not required. >>
References to sources	Article 4 (3) Aliens Act, first sentence

Table 13: Fact frame for determining whether a 'granting a temporary regular residence permit under restrictions'

The determining of the period of validity of the residence permit is done in a comparable way. The rules for determining the period of validity of a regular residence permit are laid down in the Aliens Regulation, a regulation that contains regulations that are created by the Minister of Justice and Security pursuant to an order in council made on the basis of an act as described in the Aliens Act.

The last remaining duty is the duty provide the alien with a document proving lawful residence. There are five types of documents, document I is the type of document that is given to aliens with a temporary regular residence permit. The permit contains personal data of the alien, the restriction under which the permit is valid, and the period of validity.

5 Results and Conclusion

This paper gives an overview of the application of a method for the interpretation of sources of norms to interpret the Aliens Act, and to make administrative specifications of tasks that are the responsibility of the IND. A first validation of these results has been made by a lawyer and a knowledge engineer working for the IND. In the next months there will be more extensive validations. The IND is about to decide whether to permanently use the presented method for making normative specifications.

The IND is about to decide what method to use to make traceable specifications for all tasks, products and processes administrated by the IND. Suitable alternative methods that can be used on industrial scale have not been found yet.

Results on the mapping of normative concepts on the information architecture are only preliminary. However, the FLINT representation is well received by experts in all relevant disciplines because:

1. They are perceived not to be merely technical but based on sources of norms.
2. The extensive references to sources provides insight in relations between information concepts, that cannot be (easily) found within the current information system systems, nor their architecture.
3. The modular approach, starting with a high-level interpretation that enables validating comprehensiveness, combined with the possibility to go into any level of detail, where necessary.

Further debate and validation of the results presented in this paper, may result in changes in the interpretation. It is one of the purposes of this method to clarify chosen interpretations and support argumentation about what is the 'right' interpretation [2][3]. We will report on the outcome of this process, and on possible modifications of the method resulting from it.

The analysis of the entire Aliens Act was performed by one single person within one month. This shows that a high-level analysis of the sources relevant to a complex governmental agency could be done in a relatively short period of time. Discussions on priorities for the elaboration of the results are taking place. Large-scale application of the method, working with multiple analyzers in multiple organizations demand further development particularly the development of tool support for modelers as well as the generation of IT-components. Both activities have been planned and first results are expected within the current year. In the coming weeks the IND will set priorities for elaborating details in preconditions. Furthermore, comparable interpretations of the Dutch Nationality Act, GALA and the General Data Protection Regulation (GDPR) will be made.

ACKNOWLEDGMENTS

The Dutch Science organization NWO for sponsoring projects like the AGILE project, DL4LD, VWData and others that allow us to work on normative systems, which we believe are pivotal for creating the responsible AI systems of the future.

REFERENCES

[1] Robert van Doesburg, Tijs van der Storm and Tom M. van Engers. 2016 CALCULEMUS: Towards a Formal Language for the Interpretation of Normative Systems. In: *AI4J Workshop at ECAI 2016*, The Hague, Netherlands.

[2] Robert van Doesburg and Tom M. van Engers. Using Formal Interpretations of Legal Sources for Comparing the Application of Exclusion Clauses of the UN Refugee Convention. In: *Jusletter IT* (Feb. 2018), 175-184.

[3] Robert van Doesburg, Tom M. van Engers. 2018. Arguments on the Interpretation of Sources of Law. In: AI Approaches to the Complexity of Legal Systems. AICOL 2015, AICOL 2016, AICOL 2016, AICOL 2017, AICOL 2017. Lecture Notes in Computer Science, vol 10791. Springer, Cham, 487-492. DOI: https://doi.org/10.1007/978-3-030-00178-0_33

[4] Maria Dymitruk, Réka Markovich, Rūta Liepiņa, et al. 2018. Research in Progress: Report on the ICAIL 2017 Doctoral Consortium. Artif. Intell. Law 26, 1 (March 2018), 49-97. DOI: https://doi.org/10.1007/s10506-018-9220-6

[5] Tom M. van Engers and Sjir Nijssen. 2014. From legislation towards the provision of services. In: Electronic Government and the Information Systems Perspective Lecture Notes in Computer Science, pp. 163-172. DOI:10.1007/978-3-319-10178-1_13.

[6] European Union. 2015. Joint Practical Guide of the European Parliament, the Council and the Commission for persons involved in the drafting of European Union legislation. DOI: https://doi.org/10.2880/5575

[7] Yiwei Gong. 2012. *Engineering Flexible and Agile Services: A Reference Architecture for Administrative Processes.* Dissertation. Delft University of Technology, Delft, Netherlands.

[8] Dutch Government. Aliens Act (in Dutch). https://wetten.overheid.nl/BWBR0011823/.

[9] Dutch Government. General Administrative Law Act (in Dutch). https://wetten.overheid.nl/BWBR0011823/

[10] Hennin Herrestad. 1991. Norms and Formalization. In *Proceedings of the 3th International Conference on Artificial Intelligence and Law (ICAIL '93).* ACM Press, New York, NY, 175-184. DOI: https://doi.org/10.1145/112646.112667

[11] Wesley N. Hohfeld. 1913. Some Fundamental Legal Conceptions as Applied in Judicial Reasoning. *Yale Law Journal* 23(1), 16-59.

[12] Albert Kocourek. 1930. *An Introduction to the Science of Law*, Little, Brown, and Company, Boston.

[13] Patrice Kordelaar, Freek van Teesling and Edwin Hoogland. 2010. *Acquiring and Modelling Legal Knowledge Using Patterns: An Application for the Dutch Immigration and Naturalisation Service.* In: *Knowledge Engineering and Management by the Masses (EKAW 2010).* Lecture Notes in Computer Science, vol 6317. Springer, Berlin, Heidelberg.

[14] Mariette Lokin. 2018. *Wendbaar Wetgeven* (Agile Law Making), PhD thesis, Boom Juridisch, The Hague, Netherlands. (in Dutch)

[15] Emile de Maat. 2012. *Making Sense of Legal Texts.* PhD Dissertation. University of Amsterdam, Amsterdam, Netherlands. SIKS dissertation series no. 12-26

[16] Giovanni Sileno. 2016. *Aligning Law and Action: a conceptual and computational inquiry.* Ph.D. Dissertation. University of Amsterdam, Amsterdam, Netherlands. SIKS Dissertation Series No. 2016-37.

Automated Narrative Extraction from Administrative Records

Karine Megerdoomian†
MITRE Corporation
McLean, VA, USA
karine@mitre.org

Karl Branting
MITRE Corporation
McLean, VA, USA
lbranting@mitre.org

Charles E. Horowitz
MITRE Corporation
McLean, VA, USA
chorowitz@mitre.org

Amy B. Marsh
MITRE Corporation
McLean, VA, USA
amarsh@mitre.org

Stacy J. Petersen†
MITRE Corporation
McLean, VA, USA
spetersen@mitre.org

Eric O. Scott
MITRE Corporation
McLean, VA, USA
escott@mitre.org

ABSTRACT

The U.S. Probation and Pretrial Services Office staff produce billions of pages of information on defendants' and offenders' profile and conduct. While it is critical for probation officers and district chiefs to have up-to-date knowledge on their clients to better assist and reduce risk of recidivism, the data are often stored in narrative texts in multiple large documents. As a result, these records remain mostly out of reach without the use of painstaking manual review. This paper describes an analytic prototype developed to automatically acquire structured information from natural language text in probation office documents through the application of PDF content extraction, text mining, and language analytics. Since serious mental illness is very prevalent in the U.S. corrections system, the first phase of the project focused on extracting information and constructing timelines from narrative text regarding the defendants' mental health conditions, substance use and treatment history.

Automated narrative extraction and the construction of an event timeline for defendants' mental and emotional health history have allowed the probation office to have a better understanding of their client population and to perform analyses that were previously unavailable to the organization. This technical approach can be applied across organizations, clinical administrations, and government agencies that maintain large amounts of information in the form of free text narratives.

Automated Narrative Extraction from Administrative Records*

Karine Megerdoomian
MITRE Corporation
McLean, VA, USA
karine@mitre.org

Karl Branting
MITRE Corporation
McLean, VA, USA
lbranting@mitre.org

Charles E. Horowitz
MITRE Corporation
McLean, VA, USA
chorowitz@mitre.org

Amy B. Marsh
MITRE Corporation
McLean, VA, USA
amarsh@mitre.org

Nick Modly
MITRE Corporation
McLean, VA, USA
nmodly@mitre.org

Stacy J. Petersen
MITRE Corporation
McLean, VA, USA
spetersen@mitre.org

Eric O. Scott
MITRE Corporation
McLean, VA, USA
escott@mitre.org

Sujit B. Wariyar
MITRE Corporation
McLean, VA, USA
swariyar@mitre.org

ABSTRACT

The U.S. Probation and Pretrial Services Office staff produce billions of pages of information on defendants' and offenders' profile and conduct. While it is critical for probation officers and district chiefs to have up-to-date knowledge on their clients to better assist and reduce risk of recidivism, the data are often stored in narrative texts in multiple large documents. As a result, these records remain mostly out of reach without the use of painstaking manual review. This paper describes an analytic prototype developed to automatically acquire structured information from natural language text in probation office documents through the application of PDF content extraction, text mining, and language analytics. Since serious mental illness is very prevalent in the U.S. corrections system, the first phase of the project focused on extracting information and constructing timelines from narrative text regarding the defendants' mental health conditions, substance use and treatment history.

Automated narrative extraction and the construction of an event timeline for defendants' mental and emotional health history have allowed the probation office to have a better understanding of their client population and to perform analyses that were previously unavailable to the organization. This technical approach can be applied across organizations, legal institutions, clinical administrations, and government agencies that maintain large amounts of information in the form of free text narratives.

CCS CONCEPTS

• Artificial intelligence • Information extraction • Natural language processing • Semantics • Machine learning • Temporal reasoning

KEYWORDS

Narrative analytics, Text mining, Timeline generation, Knowledge discovery, Natural language processing, Document processing, Graph representation, Syntax

ACM Reference format:

Karine Megerdoomian, Karl Branting, Charles Horowitz, Amy Marsh, Nick Modly, Stacy Petersen, Eric Scott and Sujit Wariyar. 2019. Automated Narrative Extraction from Administrative Records. In *Proceedings of the Workshop on Artificial Intelligence and the Administrative State (AIAS'19).*

1 Introduction

The U.S. Probation and Pretrial Services Office (PPSO) staff supervise more than 300,000 people a year and collect and produce billions of pages of information on defendants' and

*Throughout this document, all names of people, places, facilities and dates are replaced with fictitious ones to anonymize the information.

AIAS'19, June, 2019, Montreal, Quebec Canada

offenders' profile and conduct, as well as on the strategies and actions of officers and their outcomes. While it is critical for probation officers to have up-to-date knowledge on their clients to reduce the risk of recidivism, the data are often stored in narrative texts in multiple large documents, making it very challenging and time-consuming to collect all relevant case information manually. This renders 70 terabytes of mostly unstructured data on more than a million defendants, and strategies used by thousands of officers over decades, mostly unusable by PPSO [1]. As a result, policy makers, program evaluators, and probation and pretrial services staff have been denied valuable data with which to do their jobs.

A significant number of offenders supervised by the U.S. probation services have a current mental health condition, most of them with co-occurring substance use disorders. Defendants who suffer from mental disorders often require more intensive monitoring and specialized treatment [2]. We therefore focus on addressing important PPSO business questions to better understand the nature of the mental conditions in the officers' caseload and gain knowledge of the defendants' diagnosis and treatment history. The information was automatically obtained from the free text sections of Presentence Investigation Reports (PSIR), which represent investigations into the history of the person convicted of a crime before sentencing to determine if there are extenuating circumstances. To automatically extract and analyze the free text information in the PSIRs, we applied language analytics technology to detect the events of interest (substance use, diagnosis, treatment sessions, prescriptions) in the defendant's life and visualized them as a timeline of activities that could be reviewed by the probation and parole officers.

The system leverages Apache cTAKES (clinical Text Analysis and Knowledge Extraction System), an open-source Natural Language Processing (NLP) system developed specifically to extract and analyze clinical information from unstructured text [3]. cTAKES identifies clinical terms such as drugs, diseases and disorders, symptoms, and medical and treatment procedures. It also performs deep textual analysis and can identify, for instance, if a sentence is negated or not, or if the person being discussed is the patient or a family member. The prototype system combines the results of cTAKES with rich linguistic analysis from other open source systems such as concept ontologies and the Stanford CoreNLP parser and entity recognizer [4]. These syntactic and semantic analyses are then enhanced to adapt to the use case, by identifying significant terms for the events of interest for the mental health domain, applying linguistic analysis to improve argument and negation detection, and implementing recent advances in NLP to improve precision (e.g., vector space semantics, algorithms for building a narrative timeline).

All extracted information on a defendant's narrative is stored in a graph database and displayed on a dynamic map, allowing filtering of results based on judicial district, defendants' demographic information (age, education, citizenship), criminal category, mental conditions or medications prescribed.

As large amounts of information in business, government and administration are maintained in the form of narratives (clinical records, legal and financial summaries, progress reports, human resources assessments, etc.), the approach described in this paper for acquiring structured information from narrative text can be reapplied across organizations and government agencies.

2 Background

Past clinical information extraction systems have tended to rely on shallow NLP techniques (pattern-matching, simple parses, linear pattern interpretation rules). More recently, however, several projects have adopted knowledge-based approaches adapted for the clinical domain.

While the advantages of machine learning methods for information extraction cannot be denied, they also present a number of limitations in applications for narrative extraction from clinical data. To begin with, machine learning algorithms require large amounts of training data which are pre-tagged for the relevant features and parameters. Preparing the pre-annotated data sets can be time-consuming and expensive. In addition, such probabilistic approaches might miss rare phenomena that need to be identified since they do not occur often enough in the training data to be picked up by the learning algorithms. Another challenge for using machine learning methods in the clinical domain is that users often expect high level of consistency in the results and precise information on how the computational decisions were made. In such instances, a rule-based approach might be more transparent and easier to understand and modify.

The approach described in this paper leverages in-depth linguistic and semantic analysis to detect the domain information in narrative text, more in line with recent knowledge-based approaches [5] [6]. Machine learning approaches often require a large amount of pre-annotated data on which to train the algorithms. Since the PSIR data had not previously been tagged for the events of interest and mental conditions, a purely machine learning approach was not readily available. Hence, the prototype applies a hybrid method. It leverages rich linguistic and semantic information through the application of open-source Natural Language Processing systems, adapted for the existing use case by applying a combination of rule-based linguistic analysis, vector space semantics, and machine learning techniques to enhance the

results. These were used to improve negation detection and argument identification (i.e., entities the events refer to), and to develop temporal reasoning algorithms. Ontologies (lexicons) of mental health and medication terms, vetted by a subject matter expert, were used for concept identification. The rest of this section provides a detailed description of the technical steps in building the analytic prototype.

3 Technical Approach

The technical approach is a hybrid one, leveraging open source NLP applications often developed by training machine learning algorithms, and refining the syntactic and semantic analyses with a combination of knowledge-based and probabilistic approaches.

3.1 Analytic Pipeline

The presentence reports undergo several steps in order to extract the defendant's mental health and substance use narratives. These are shown in Figure 1 and are described in detail in the rest of this section. The specific steps involved are:

1. **Content Extraction:** parsing the different sections of the PDF documents and extracting the structured profile and criminal information as well as all free text content. This component also "cleans" the data by normalizing the textual content to maximize processing.
2. **Language Analytics:** The extracted text for each PSIR is run through the Natural Language Processing components, providing a full linguistic parse, a list of entities and events of interest, and semantic relationships.
3. **Knowledge Discovery:** This step is the heart of the textual analytics where the system identifies all concepts, events, and their relationships for the domain of interest.
   - Identifies the events of interest associated with the defendant (arrests, diagnoses, treatments, prescriptions, drug use, suffering from a mental condition);
   - Determines whether the information is obtained from medical records or if it is reported by the defendant, by a medical professional, or by a third party;
   - Provides full event description including date, location, persons involved, treatment provider, nature of treatment and medication prescribed;
   - Computes the temporal relationships between the various events to build a narrative timeline for a defendant.
4. **Neo4j Database:** Neo4j is a graph database management system and is available as open source software. All extracted information from the Knowledge Discovery component, as well as the client demographic metadata, and structured information on arrest history and federal offenses extracted from the presentence reports are loaded into the database.
5. **User Interface (UI):** This component interacts with the Neo4j database and displays results on a Google Earth map. The UI allows the user to run queries, to review the details on particular defendants, and to see aggregate results on the data set.

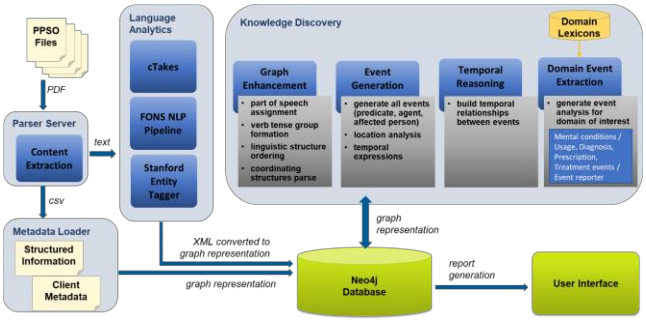

Figure 1: Analytic pipeline for narrative extraction and timeline development

3.2 Content Extraction

The Content Extraction component parses the PDF presentence reports, identifies all subsections and extracts the textual content. To analyze the mental health and substance use information of defendants, the text content of the Mental and Emotional Health (MEH) and Substance Abuse (SA) sections in presentence reports are automatically extracted. In addition, this step identifies and extracts all federal charges from the cover sheet of the PSIR, criminal history information from the Juvenile Adjudications and Adult Criminal Convictions sections of the report, Arrest Dates and associated charges from the Criminal History information, and Criminal History Score and Category from the Criminal History Computation section.

The prototype's Content Extraction component successfully extracted information from 92% of the original PDF documents, providing us with a data set of 11,243 extracted narrative text documents to analyze. Given that some defendants have more than one presentence report associated with them, the successfully extracted content corresponds to 10,973 defendants. The free text content extracted from the MEH and SA sections amount to 22,486 text items. These can range from a few sentences to several paragraphs depending on the report.

3.3 Language Analytics

The Language Analytics component leverages existing Natural Language Processing software to perform various linguistic analyses on a piece of text. NLP is a subset of Artificial Intelligence (AI) and is fast becoming an essential technology in modern-day organizations to gain significant insights from unstructured content, such as email communications, social media, videos, customer reviews, customer support request, and administrative records in business and government. Natural Language Processing tools and techniques help to

automatically process, analyze, and understand large amounts of data, providing structure and meaning to information that originally was in unstructured form.

In this step of the analysis, the texts extracted from the Mental and Emotional Health and Substance Abuse sections of the PSIRs are run through several NLP software tools. The software packages currently in use are Apache cTAKES (clinical Text Analysis and Knowledge Extraction System), Stanford Named Entity Recognizer, and FONS (Framework for Operation NLP Services)[1].

cTAKES output forms the primary basis for further analytics. It was chosen primarily because of its entity recognition capabilities in the clinical domain, which aligned with the desire to obtain data about PPSO clients' mental and emotional health and substance use. Entities identified by cTAKES include medical conditions, drugs/medications, medical procedures, and medical symptoms. The entities identified by cTAKES out-of-the-box were supplemented with additional entities frequently encountered by analysts in PSIRs. We worked closely with a PPSO subject matter expert to review the list of conditions and medications that cTAKES recognized, and identify the ones that were of interest in the mental and emotional health and substance use domain. The subject matter expert also identified a more general superclass for each of these specific mental and emotional conditions so that further analysis could be conducted at the appropriate level of granularity. For example, conditions such as *depression, chronic depression,* and *major depressive disorder* were all mapped to the more general term *depressive disorder.*

cTAKES also provides domain-independent NLP capabilities of syntactic parsing, dependency parsing, and semantic role labelling – it can give the base forms of words, their parts of speech, mark up the structure of sentences in terms of phrases and syntactic relations, detect negation in the sentence and identify the role of the entities in a sentence (e.g., agent of event). The results of all these capabilities were used to identify events of interest in a client's mental and emotional health and substance use history. However, we found it useful to supplement the cTAKES output with other natural language processing systems to achieve the most accurate analysis. The Stanford Named Entity Recognizer was applied to identify people, places, organizations, dates, times, and locations, none of which are identified by cTAKES. Additionally, the FONS system, which also generates entities, syntactic parsing and dependency parsing output, was used to supplement cTAKES' output to obtain a higher level of accuracy. In particular, FONS was applied to the PSIR text data to tag entities (people, facilities, locations, dates and times), and to categorize all events into conceptual classes by detecting event types (e.g., state, transfer, communication) and different verb meanings (e.g., *prescribe* can either be the verb denoting the prescription of medication by a medical professional or a communication event meaning 'to advise', 'to recommend').

3.4 Domain-Specific Entity and Event Identification

The Knowledge Discovery phase of the analytics involves processing the output from the Natural Language Processing systems to perform several steps in knowledge discovery in natural language text:

1. Identify concepts (entities and events) of interest associated with the client, including mentions of a client suffering from a mental condition, diagnoses, treatments, prescriptions and drug use.
2. Detect the event description such as the date and location when it occurred, the persons involved, the treatment provider, the nature of treatment (e.g., inpatient or outpatient, anger management, drug rehabilitation) and the medication prescribed.
3. Detect the source of the information – was the information reported by the client, was it obtained from medical records or a medical professional, or reported by a third party?

As described, cTAKES detects these entities of interest in the mental and emotional health domain. However, to identify whether a client is suffering from a mental condition, it does not suffice to simply retrieve sentences with a mental condition mention. It is also important to detect the subject of the sentence to distinguish cases where a family member is mentioned to suffer from a mental condition (e.g., "*the defendant's mother suffered from Schizophrenia"*), and to exclude any negated events (e.g., "*the defendant does not suffer from a severe mental disease or defect"*). Fortunately, when cTAKES identifies a concept, it also identifies that sentence's polarity (whether the entity appears in a negated context or not), and the event's subject (whether that event or concept should be ascribed to the client described in the text, a family member of the client, or someone else). Some modifications to the cTAKES source code were made to improve the accuracy of these attribute identifications.

While the cTAKES entities can be counted to obtain statistics on the prevalence of various mental conditions among the defendant population, further processing is necessary to identify more complicated events, such as receiving a diagnosis, attending treatment, being prescribed medication, or using drugs. To identify the events of interest, a small sample of PSIRs was reviewed to identify the verbs commonly

[1] FONS is a software package pipeline leveraging open source tools and was built by a research team at MITRE to detect events of interest to national security.

associated with these events. An iterative process was used in reviewing the event detection results and updating the predicates for the domain. The verbal predicates associated with each type of event are listed in Table 1.

Event Type	Predicate
Diagnosis	diagnose
Prescription	prescribe, treat (with)
Treatment	admit, attend, complete, discharge, enroll, enter, hospitalize, meet, participate, place, receive, see, seek, speak, treat, undergo
Usage	abuse, addict, consume, drink, experiment, ingest, inhale, relapse, smoke, snort, take, try, use

Table 1: Verbs used to identify events related to mental and emotional health and substance use

Once the predicates are identified, the semantic roles associated with each occurrence of the predicate are automatically extracted to enable the identification of the predicate's agent, affected entity, and whether the predicate was negated. The sentence in which the predicate appeared was also examined to identify medications, drugs, mental conditions, medical procedures, and treatments associated with that event.

To detect the source of the information, all sentences with Communication events identified by the FONS software package were analyzed and the subject of the verbs extracted. For example, in "*Dr. Gray stated that the defendant has never been hospitalized for emotional disorders of any kind",* the communication verb *stated* is detected and its subject, *Dr. Gray* (a medical professional), is identified as the source of the information. Similarly, in the example "*the defendant's mother also reported he was diagnosed with Bi-Polar Disorder several years ago",* the source of information is identified as *the defendant's mother* (a third party).

If the subject of the communication verb is mentioned as *the defendant*, the system treats it as a self-reported event. In the writing style of the presentence reports, mentions of *he* or *she* tend to refer overwhelmingly to the defendant. Since the current version of the analytic system does not include a "coreference resolution" component that can accurately identify who the pronouns refer to, the assumption is made to treat these cases as self-reported events. This can be seen in the following example where the events in both sentences are automatically labeled as self-reported: *"The defendant expressed feelings of depression, helplessness, and hopelessness. He also admitted to occasional auditory hallucinations."*

If the name of the defendant is mentioned as the subject of the communication verb (e.g., "*McKenna could not recall being prescribed medication to treat his Depression"*), an additional step is performed to verify the name *McKenna* against the defendant metadata information – if the system finds a match, then the information is labeled as self-reported.

Certain automated enhancements had to be made to the Communication event detection, however, since the automatic classification by FONS included verbs such as *stuttered* and *snorted*. In order to improve the results, we computed semantic vector measures that capture the similarity in usage of verbs against canonical Communication events such as *reported* and *stated*. The verbs that are closest in the context of use within the text and thus have closer meaning to the *report/state* verbs produce higher values and are thus more likely to be indicative of the source of information. The top verbs identified as Communication events are listed in Table 2.

Event Type	Predicate
Communication	state, indicate, note, explain, report, say, acknowledge, discuss, identify, confirm, deny, address, agree, communicate, question, suggest, tell, describe, claim, mention, inform, disclose
Other formulation	according to

Table 2: Terms used to identify the source of information.

This linguistically rich event-based narrative analysis methodology allows the Language Analytics component to extract information of interest including the people involved in the event, the time it occurred, and the places mentioned. A sample analyzed sentence is shown in the following example:

> The defendant<source-of-info> reported she<affected-entity/diagnose-event> was diagnosed<diagnose-event> at the age of 14<time> with depression<mental-condition>, schizophrenia<mental-condition> and bi-polar disorder<mental-condition> and was not prescribed<prescribe-event|NEGATIVE> any medication<medication-mention>.

3.5 Generalized Event Analysis

While cTAKES proved very useful for identifying events in the clinical domain, it is not specifically tuned for identifying more general events. Events that are not directly related to diagnoses, prescriptions, substance abuse, or treatment may still be of interest when analyzing a client's mental and emotional health history. For example, in *"He became depressed when his infant brother died"*, the event of the infant brother's death does not fall into one of the domain-specific event categories, but it is still relevant to indicate a trigger or risk factor. To try to capture these types of events, a more general approach to parsing free text was used, producing an event-

based analysis for every verb encountered in the Mental and Emotional Health and Substance Abuse sections.

As part of the Knowledge Discovery phase, the linguistic output from the NLP systems loaded into the Neo4j graph database is used as the basis for generating events that do not rely on a domain-specific vocabulary. In this framework, events are generally identified by the presence of a verb and an event-based analysis is performed on the sentence. In simple sentences, this means that one event corresponds to the entire sentence. However, if a sentence contains multiple clauses, each clause could potentially represent one event. In the sentence *"he became depressed **when** his infant brother died"*, *becoming depressed* is one event, and *his infant brother died* is a separate event. The two clauses are linked by the conjunction *when,* which indicates the temporal relationship between them. To handle sentences such as these, a list of terms that signify a subordinate clause was created and sentences were divided into clauses when one of these terms was found. The list of terms used is in Table 3 below. These terms are used in further analytics to identify temporal or causal relations between events.

Relationship Type	Clause Marker Terms
Temporal	after, before, during, following, prior to, throughout, until, upon, when, while
Causal	although, as a result of, because, due to, in order to, since
Other	according to, along with, in addition to, relating to

Table 3: Terms signifying the presence of a subordinate clause in a sentence.

After all clauses have been identified, an event is generated for each clause. If the clause contains a verb, the verb phrase forms the basis of the event. If there is no verb phrase in the clause, (e.g., in the sentence "w*hile in prison, the defendant used heroin",* "while in prison" is a clause without an explicit verb), the phrase after the clause marker forms an event description which is the basis of the event. Then, information from the syntactic parses, dependency parses, semantic roles, and named entities are used to identify agents, affected entities, indirect objects, locations, and temporal mentions related to the basis of the event for a complete narrative analysis.

3.6 Temporal Reasoning

Once all relevant events have been extracted from the text of the PSIRs, it is possible to make a timeline of the relevant events with temporal mentions in a client's history. To accomplish this, we adapted TimeML (Markup Language for Temporal and Event Expressions) standards to the narratives generated [6]. TimeML is designed to provide a standard way to annotate events with a time stamp and place events in chronological order; it is thus optimal for the problem of timeline generation. In TimeML, events are typically described by verbs, which aligns with our approach to narrative generation. In the actual TimeML specification, temporal expressions are marked as separate entities, falling into the categories of ***date*** (for events that take place at a specific time, which might be a date, month, or year), ***time*** (for events that take place at a specific time of day), ***duration*** (for events that have clear start and end points), and ***set*** (for periodic events). In our adaptation, we recorded the type of temporal expression as an attribute of the event it was associated with, and did not use the category of ***time*** since the specific time of day of various events is not typically specified in PSIRs. The category of ***set*** was recorded but is not currently used for timeline generation. The start date and end date of each event are also recorded as additional attributes.

To place events in order, TimeML uses the TLINK tag, which records the id's of two related events and the temporal relationship between the two. In this project, temporal relationships are marked as an attribute of the event rather than a separate entity, and an abbreviated set of seven temporal relationships are used, rather than the fourteen defined in TimeML. The temporal relationships utilized are listed in Table 4.

Relationship	Description
AFTER	Event 1 occurs some time after Event 2
BEFORE	Event 1 occurs some time before Event 2
BEGINS_AT	Event 1 occurs immediately after Event 2
ENDS_AT	Event 1 occurs immediately before Event 2
INCLUDES	Event 1 starts before and ends after Event 2
IS_INCLUDED	Event 1 starts after and ends before Event 2
SIMULTANEOUS	Event 1 and Event 2 start and end at the same time

Table 4: Temporal relationships used for timeline generation.

Determining the values of the temporal type, start time, end time, and temporal relationships for the generated events is a three-step process. In the first pass through the events, any temporal mentions associated with each event were parsed with regular expressions and used to set the event's type, start date, and end date. Next, temporal relationships were identified by examining events to see if they contained any of the subordinate clause markers listed in Table 5. Rules were then applied to relate two events connected by a subordinate clause marker. One final pass through the events was used to set any additional start and end dates that could be inferred after the temporal relationship was determined.

We can follow this entire process on the sentence "*He began smoking marijuana at the age of 16 until his arrest in 2014",* which contains the events *he began smoking marijuana* and *his arrest in 2014.* The first step after the identification of the two clauses is to identify the presence of the temporal expressions in each clause – *at the age of 16* in the *smoking* event (EV1) and *in 2014* in the *arrest* event (EV2). In EV1, the start time can be obtained from the defendant's date of birth in the profile information available in the database. In EV2, the Knowledge Discovery component establishes that the temporal expression is of type ***date***, with a start and end time set to span the whole year as shown in Table 5, since the time is not more clearly specified than that. The second step will identify the subordinate clause marker *until*, and follow a rule that establishes that the *smoking marijuana* event ended at *his arrest in 2014.* The final step will use the presence of the ENDS_AT relationship to set the end time of *he began smoking marijuana* to the start time of *his arrest in 2014*. The final event analysis associated with a temporal range is then used to build a timeline and visualize on the web-based interface.

Clause/Event detection	[He began smoking marijuana]$_{\text{clause1/EV1}}$ **until** <clause-marker> [his arrest in 2014] $_{\text{clause2/EV2}}$
Step 1: Detect temporal expressions when available	EV1: <type: "date", startAt: {year: '1992', month: '6', day: '12'}, endAt: None> EV2: <type: "date", startAt: {year: '2014', month: '1', day: '1'}, endAt: {year: '2015', month: '1', day: '1'}>
Step 2: Establish temporal relation between events	id: "EV1", relType: "ENDS_AT EV2"
Step 3: Temporal reasoning to set temporal expression	EV1: <type: "date", startAt: {year: '1992', month: '6', day: '12'}, endAt: {year: '2014', month: '1', day: '1'}> EV2: <type: "date", startAt: {year: '2014', month: '1', day: '1'}, endAt: {year: '2015', month: '1', day: '1'}>

Table 5: Temporal reasoning process in Knowledge Discovery phase.

4 Graph-Based Representation

The main motivation for using a graph database to store the parse output is that syntactic parse outputs are often modeled in linguistic theory in the form of trees (a graph in which each node has a single parent) and dependency parses capture the semantic relationship associated with two nodes, so storing the parse outputs as a graph allows to use Neo4j API (Application Programming Interface) and CQL (Cypher Query Language) to directly access these grammatical relationships and handle the recursion inherent in language. Additionally, once the natural language parsing outputs are stored in graph format, it is easy to align and merge the outputs from the different NLP systems being used. Finally, Neo4j provides a visualization of the graph for linguists and developers that assists in understanding the structure of the language.

Once the output from the NLP systems is stored in the database, we apply several enhancements to the raw system output to improve the parses' accuracy and generalizability. These enhancements include labelling all nodes with a more coarse-grained part-of-speech tag, grouping together multi-word verb phrases into a single entity (e.g. merging the nodes for the terms in the phrase *has been attending* into a single node *attend* with appropriate tense and aspect information), and combining coordinated phrases with conjunctions into a single entity (e.g. merging the nodes for the terms in the phrase *mental and emotional* into a single node to facilitate further analysis).

The User Interface interacts with the Neo4j database to access all content and narrative analytics output and displays the results on a Cesium Server. The web-based interface allows the user to run queries of interest, filter based on the defendant's profile information, and view the retrieved information on a spatial map of judicial districts or States.

The UI displays an aggregate report of the data for the provided query as shown in Figure 2. This display can be further filtered based on the mental conditions, medications and substances of interest, as well as the defendant's demographic information and criminal category.

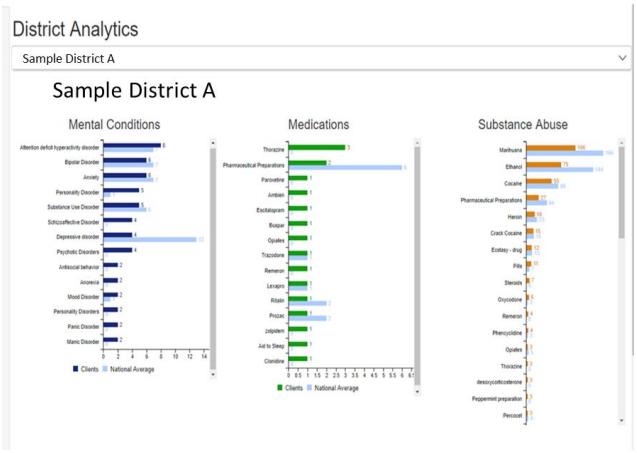

Figure 2: Narrative analytics results viewed by judicial district.

The user can then select to view the identified defendants on a map to the level of street detail. The user may also select to view a particular client's information in more detail, such as mental conditions reported, and see associated text from the Presentence Investigation Report with relevant sections highlighted. In addition, the data are used to visualize a timeline of the defendant's life events including arrests, diagnoses, substance use, and treatments.

5 Results of Analytics

The programmatically important questions of interest to PPSO that are addressed in the current prototype are (i) determining how many defendants sentenced had a mental health condition; (ii) the types of conditions present; (iii) the source of the diagnosis; (iv) prior treatment exposure; and reporting of that information by demographic, offense and prior criminal history information.

To identify the number of defendants with a mental illness, the system extracts all the client cases where a mental illness was mentioned as attributed to the defendant (whether officially diagnosed or not). It was found that 3,959 defendants in the data set (about 36% of the studied population) had a history of one or more mental conditions. If Substance Use Disorder is included as a mental health condition, that number increases to 58%. Figure 3 provides the heuristics for the mental health conditions mentioned in the Mental and Emotional Health sections of presentence reports studied. However, the total number of defendants who have officially been diagnosed with a mental condition is 2,238 (20% of the studied population). In addition, 82% of the defendants had a history of substance use (mainly Marijuana and alcohol), and 53% of cases had a prior criminal record cited. Most common prescriptions are Prozac, Ritalin, Seroquel and Xanax, and top substances include Marijuana, Alcohol, Cocaine, Methamphetamine and Heroin.

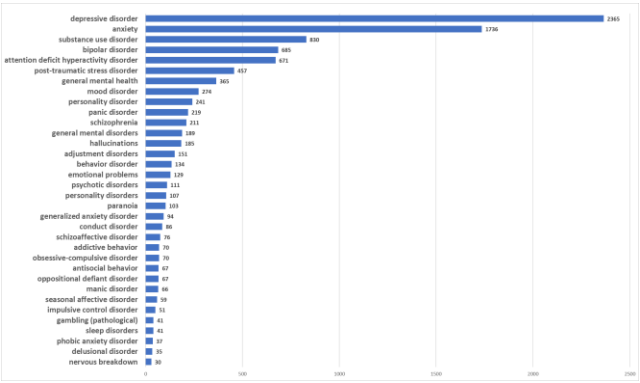

Figure 3: Mental health conditions associated with about 11,000 defendants.

As described earlier, the analytic prototype identifies the source of information for each detected event of interest. There are five distinct categories for the source: (i) self (client self-report), (ii) medical professional, (iii) medical records, (iv) report (official non-medical records, including evaluations and assessments), and (v) third party (third party corroboration such as a family member, defense counsel, probation agent, or pretrial services agency). In the presentence reports studied, the majority of the events (about 89% of all events found) are self-reported.

The full set of results in response to the PPSO business questions is shown in Table 6.

Category	Count	Percent	Accuracy*
PSRs with Mental Health and Substance Abuse sections [set P]	11,243		
Total number of clients corresponding to set P	10,973	100%	
Total number of clients for which an event of interest was found (diagnosis, prescription, treatment, substance use, suffering from mental condition)	10,743	98%	
Total number of clients with a mental condition	3,959	36%	98%
Total number of clients who had a diagnosis made, with an explicit mental condition	2,238	20%	99%
Total number of clients who attended some sort of treatment or assessment/evaluation (mental health)	3,469	32%	87%
Total number of clients who attended some sort of treatment or assessment/evaluation (substance abuse)	3,817	35%	90%
Total number of clients with medication prescribed	2,057	19%	92%
Total number of clients using/abusing substances	8,998	82%	*Not Evaluated*
Total number of clients with historical arrest information (30,566 arrests total, extracted from metadata)	5,764	53%	N/A

Table 6: Automatically obtained responses to the PPSO business questions on defendants' mental health and substance use history.

System performance was evaluated by creating a small reference sample of about 500 sentences to measure the accuracy of the information extracted for each event type. The 500 sentences were manually annotated by team members indicating the expected mental conditions, event types (diagnosis, treatment, prescription, usage), and medications. The annotations also included important event-related information such as the agent (prescriber, diagnoser), polarity (whether the event is negated or not), and the temporal expression associated with the event. The language analytics results were then compared to the pre-annotated reference set to measure how many of the detected elements were accurate and to also calculate how many of the expected elements were not picked up by the system.

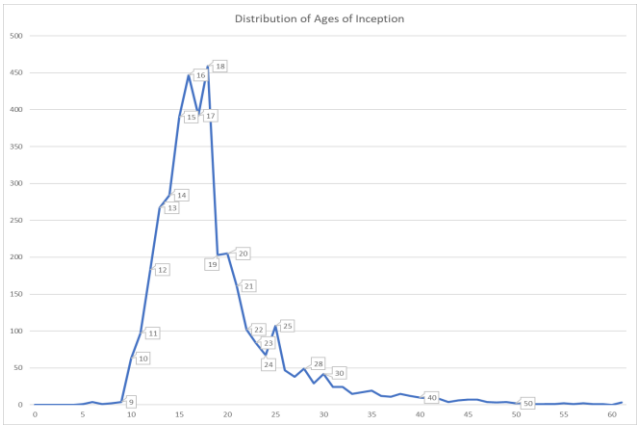

Figure 4: Correlation analysis shows defendants' onset of substance use. The x-axis represents the defendant's age and the y-axis is the number of times the onset of substance consumption is found in the text.

We also explored the aggregated national and district data for potential correlations and analyses across defendants. Figure 4 illustrates one such analysis, which shows the onset of substance use among the defendants studied. This examination

automatically detects any mentions of the age of the defendants in the Substance Abuse texts and identifies any sentences that refer to the onset of using a drug or alcohol by the defendant. For instance, a sentence such as "*he began using cocaine at age 17*", is labeled as an "inception predicate" and associated with the age of the defendant (i.e., 17). The results show that the onset of substance use among defendants starts at age 10, with a steady increase to age 16 and peaks at age 18.

This analysis is only one example of the types of aggregated correlations and computations that are available after the full language analytics have been performed on the data. Other correlations explored include automatically detecting instances of co-morbidity to understand which mental conditions tend to co-occur most often among the population, automatic detection of defendants with previous suicide attempts or history of suicidal ideation, and identification of events that may trigger mental health issues (e.g., death of a family member, history of sexual or domestic abuse, fatal medical diagnosis, divorce).

6 Application to Risk Assessment

Analysts working in the Probation and Pretrial Services domain leverage a variety of data-driven instruments to measure trends, train officers, and assess the recidivism risk in individual clients. At a high level, these efforts are typically described in terms of the popular Risk, Needs, and Responsivity model, which dictates that effective offender supervision ought to allocate more treatment resources to high-risk clients, that treatment should target specific criminogenic needs in the client's case, and that officers should apply cognitive-behavioral techniques to respond to the details of a client's particular situation [7, 8]. In recent years there has been a trend toward using data-driven approaches for the first step, and actuarial risk assessment instruments such as "Levels of Service" surveys [9] and the federally developed Post-Conviction Risk Assessment (PCRA) [10] have played an increasingly important role in the allocation of treatment resources. These tools are typically based on survey questions that must be administered and recorded by the officer, which then serve as inputs to traditional statistical modeling techniques (ex. linear regression). Such tools are time-consuming to use, and they offer only a limited, static snapshot of the specific criminogenic needs that are present in a client's case. Risk/needs assessment is an active area of research, and efforts are ongoing to identify next generation tools that can offer improved data-driven methods that can help support probation officer responses during their regular interactions with clients. Leveraging the wealth of unstructured information that is present in the existing documentation that is available in probation case tracking systems is one promising approach to solving this problem.

Any application of AI or data analysis to officer decision-making can end up having a significant impact on the population under supervision, and so it is important to be aware of the various ethical concerns that surround the application of data analysis software to social issues [11]. Such concerns include the need for general algorithmic accountability [12], the need for assurance that algorithms that are used for such important tasks as recidivism prediction do not exhibit unacceptable biases [13], the need for judicial review of algorithm-assisted decision-making (where such review may be called for), and more practically, the need to inspire trust in users, who tend to be unwilling to rely on algorithms whose inner workings are poorly understood. Some of these issues are of greater concern than others in a probation domain. Judicial review, for example, is a legal necessity when algorithms directly impact a judge's decisions, but risk/needs assessments for offenders who are on supervised release are not normally referred to in judicial decision-making.

In this work, we focus on the foundational question of extracting information from unstructured text that can inform the decisions of officers and analysts working within the federal probation system. We defer questions about automated risk assessment, and the fairness thereof, to future research. The current work focuses instead on extracting and arranging raw facts from various sources in a visualization that a human can use to support their professional judgement in a particular case and that an officer can potentially leverage in detecting patterns that had previously been unavailable.

7 Future Directions

The paper describes a successful approach to the automatic extraction and analysis of narrative text in the mental health and substance use domain. The approach has since been applied to other domains such as employment history and financial history. The results provide evidence that the use of technology in identifying important information in free narrative text in administrative records is feasible and cost-effective, and any adaptations to new domains can be accelerated through probabilistic methods. These analytics can be further developed in various directions, depending on the mission needs of the organization. This section provides some directions to pursue.

The current results of the analytics can further be improved upon by annotating more data and performing a larger-scale evaluation and refinement cycle. Although event extraction accuracy ranged in the 90-percentile, an evaluation conducted on a larger data set will provide better accuracy measures and can identify low frequency events that may have been missed in the current version of the analytics. Further work can also be

performed on negation and argument detection to achieve higher precision. In addition, the analytics results have not yet been fully validated by a subject matter expert to ensure that the data identified and the way the results are presented are valuable for the PPSO officer or mental health analyst.

Building a timeline of a defendant's life events from narrative text is a very complex task and the topic of much current research in the field of NLP. We successfully identified the temporal expressions associated with events and introduced a temporal reasoning component which is tightly integrated into the system's syntactic parse and semantic relations output. Yet, identifying the temporal relations between events is not an easy task and oftentimes, the system needs to infer a relationship that is not overtly mentioned in the sentence. Building a hybrid method combining knowledge-based linguistic analysis with a statistical machine learning approach will provide more robust temporal relationship analyses.

One of the issues that were left unaddressed in the current version of the analytics was the distinction between events (e.g., diagnoses, treatments) that occurred in the past and those that are currently valid. This can be accomplished by leveraging the tense and aspect information that the system computes and adding a filter on the UI to allow the user to view only events that are current.

Building a complete timeline of a defendant's life events will provide the important information at the individual level for PPSO officers to view and analyze, helping them identify precursor events and triggering factors. For instance, in addition to the mental health and substance use information, the personal history of the defendant (e.g., whether he or she graduated high school, history of domestic violence or neglect), existence of dependents (e.g., number of dependents and their age, learning issues, custodian), family relations (e.g., siblings and whether they have a criminal or substance abuse history), employment status, gang or terrorism activity, etc. are all important information elements that could shed light on the defendant's situation and allow probation officers to provide more efficient supervision and intervention measures to reduce recidivism. This requires fusing all events and information extracted from presentence documents onto a single timeline to view and analyze.

An important goal for analytics research is to leverage the large amount of data from diverse sources available to the probation office—including treatment reports, Chrono notes, social media, structured metadata, risk assessments, and court documents—to obtain a more complete picture of the defendant's history, conduct and status. The data analytics methods will be applied to these data sources and all results combined into a unified database available for query and analysis. Building on a multi-source analysis, the system can begin identifying precursor events to criminal activity or noncompliance, or detecting triggers for mental health issues or substance use relapses, and leverage that information to build a predictive model to forecast potential risk and generate automatic alerts. Such an alerting system can help direct an officer's attention to elements of a client's case history that indicate a special cause for concern.

ACKNOWLEDGMENTS

We would like to acknowledge the guidance and support of the U.S. Probation and Pretrial Services Office throughout this project. In particular, we would like to thank Steve Levinsohn for providing essential subject matter expertise to the team. The project was led and funded by the Technology Solutions Office at the Administrative Office of the U.S. Courts as part of the Applied Technology Research and Development program managed by the Judiciary Engineering and Modernization Center operated by the MITRE Corporation.

REFERENCES

[1] Matthew G. Rowland (2018). Federal Probation and Pretrial Services: What's Going On and Where Are We Going? (Presentation by the Chief of PPSO). Probation and Pretrial Services Office, Administrative Office of the U.S. Courts.

[2] Probation and Pretrial Services Office (2016). Overview of Probation and Supervised Release Conditions. Administrative Office of the United States Courts.

[3] Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. *Journal of American Medical Informatics Association* 17: 507–513.

[4] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60

[5] Shervin Malmasi, Nicolae L. Sandor, Naoshi Hosomura, Matt Goldberg, Stephen Skentzos, and Alexander Turchin (2017). Canary: An NLP Platform for Clinicians and Researchers. *Applied Clinical Informatics* 08(02): 447-453.

[6] Hyuckchul Jung, James Allen, Nate Blaylock, Will de Beaumont, Lucian Galescu, and Mary Swift (2011). Building timelines from narrative clinical records: Initial results based on deep natural language understanding. *Proceedings of BioNLP 2011 Workshop*, pp. 146-54. Association for Computational Linguistics.

[7] Don A. Andrews, James Bonta, and Robert D. Hoge (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal justice and Behavior* 17.1: 19-52.

[8] Bonta, James, and Donald A. Andrews (2007). Risk-need-responsivity model for offender assessment and rehabilitation. *Rehabilitation* 6.1: 1-22.

[9] J. Stephen Wormith and James Bonta (2018). The Level of Service (LS) Instruments. In *Handbook of Recidivism Risk/Needs Assessment Tools*, pp. 117-145. Wiley & Sons.

[10] Christoper T. Lowenkamp, James L. Johnson, Alexander M. Holsinger, Scott W. VanBenschoten, and Charles R. Robinson (2013). The federal Post Conviction Risk Assessment (PCRA): A construction and validation study. *Psychological Services* 10(1): 87-96.

[11] Scott W. VanBenschoten (2008). Risk/needs assessment: Is this the best we can do. *Federal Probation* 72: 38.

[12] Nicholas Diakopoulos (2014). *Algorithmic Accountability Reporting: On the Investigation of Black Boxes*. Columbia University, Tow Center for Digital Journalism. New York: Columbia University Academic Commons.

[13] Alexandra Chouldechova (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big Data(5): 153-163.