Induction of Narrative Models for Legal Case Elicitation

Karl Branting[1,*], Sarah McLeod[2], Bryant Park[3] and Karine Megerdoomian[4]

[1]The MITRE Corporation, McLean,VA, USA
[2]The MITRE Corporation, Seattle, WA, USA
[3]Cornell University, Ithaca, NY, USA
[4]The MITRE Corporation, Miami, FL, USA

Abstract

This paper proposes a new computational architecture for narrative-driven case elicitation, describes six new legal narrative corpora, and evaluates two different approaches to creating legal narrative schemas, the first using language models, and the second using event sequence alignment. An experimental evaluation suggests that the sequence alignment approach may be more appropriate for legal corpora that are small, sparse, and heterogeneous.

Keywords

narrative schema induction, law, computational linguistics, machine-learning, human-computer interface, event sequence alignment

1. Introduction

Increasing numbers of litigants worldwide face the challenges of representing themselves in courts and other decision forums without the assistance of an attorney. [1] [2]. Significant reductions in public legal-aid expenditures in many jurisdictions have fueled this trend, leading to increases in self-represented litigants (SRLs) in the UK [3], the EU [4], Canada [5] [6], and the United States [7]. SRLs are typically at a significant disadvantage in legal proceedings compared to parties represented by an attorney [8].

Many technological innovations, such as Online Dispute Resolution [9] [10], can assist SRLs in asserting rights, claims, or defenses, but the most widespread form of computer assistance consists of legal form-filling software [11]. A key limitation of legal form-filling software systems is that they seldom provide users any assistance in formulating narrative statements of facts. Typically, such systems are built around hard-wired decision logic in which the case information is elicited in the form of feature-value pairs, e.g., dates, dollar amounts, names, addresses, etc. The display order of the windows and data fields is often conditioned on values provided by the user via either a precalculated set of decision paths (the most common approach in current systems) [12] or through a goal-driven dynamic process based on logic-programming [13]. While some systems are capable of instantiated narrative templates with user-provided facts, none are able to interpret text provided by a user or assist a user in paraphrasing facts in a manner likely to communicate them most effectively to a judge.

Legal aid attorneys often elicit case facts by starting with a general question (e.g., "How can I help you today?") followed by a series of follow-up questions to fill-in missing parts of the client's story while ignoring the irrelevant details. Eliciting the overall story permits an attorney to reason about how well the facts fit the requirements for various legal remedies that might satisfy the client's goals and to summarize the facts in the narrative fields of petitions or other court documents.

Attorneys' narrative elicitation process is structured around their expectations about what constitutes a legally relevant story. Such expectations probably arise from hearing similar stories from numerous clients. We surmise that an automating process for narrative-driven case elicitation must, in a similar way, be based on generalizations of multiple relevant prior stories.

This paper proposes a computational architecture for narrative-driven case elicitation and describes a series of experiments in induction of narrative schemata from corpora of legal narratives. These experiments are informed by prior work on narrative schema induction but reveal distinctive challenges and constraints imposed by legal narratives.

The next section describes related work on the role of narrative understanding in legal problem solving and on narrative schema induction. Section 3 describes an architecture for narrative-driven case elicitation that comprises on offline component, in which narrative schemata are induced from corpora, and an online component, in which schemata are used for mixed-initiative dialogue. Section 4 overviews the process of converting text to event sequences, and six corpora of legal narratives of

Proceedings of the Sixth Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2023), June 23, 2023, Braga, Portugal
*Corresponding author.
lbranting@mitre.org (K. Branting); smcleod@mitre.org (S. McLeod); blp73@cornell.edu (B. Park); karine@mitre.org (K. Megerdoomian)
0000-0002-9362-495X (K. Branting)

representative types are described in Section 5. Section 6 sets forth two approaches to narrative schemata induction—one based on language models and a second based on sequence alignment—and presents experimental results in predicting missing narrative events. The implications of these results and proposals for future work are presented in the final section.

2. Related Work

Our research in creating and using legal narrative schemas for fact elicitations connects two complementary strands of prior work: investigations of the role of story understanding legal client interviewing; and induction of narrative schemas to support automated story understanding.

2.1. The role of story understanding in case elicitation

The facts of legal cases are more than mere collections of events. Rather, case facts are narratives having settings, characters with goals and motives, and events linked by temporal, causal, and intentional relations. Outcomes of trials often depend on the relative story-telling ability of attorneys and witnesses [14], and jurors have been shown empirically to decide cases based on which of two competing narratives imposes the highest degree of coherence on the evidence presented at trial [15] [16].

When interviewing a client to determine the client's story, attorneys often try to elicit "the causal and temporal connections that contribute to giving the events contextual meanings … with the aim of defining 'Who has done what, how, when, why and where?'" [17]. Clients' narratives are often "redefined to be a legally relevant narrative" by legal aid attorneys, a process that can sometimes interfere with or prevent understanding of emotionally salient background information if it is too rigid [18].

Notwithstanding the central role of narrative in legal client interviews, there has been little exploration of techniques for automating narrative elicitation. A paucity of operational theories of text-based narrative analysis may have played a role in the lack of activity in this area. Recent advances in narrative schema induction have enabled the novel research described below in this paper on narrative case elicitation.

2.2. Narrative Schema Induction

The importance of narrative schemas for story understanding was recognized early in the history of AI. Roger Schank and Robert Abelson coined the term "scripts" to indicate stereotypical sequences of events that create expectations and fill in missing details [19].

Induction of narrative schemas (i.e., scripts in Schank/Abelson terminology) was pioneered by Chambers and Jurafsky [20] who defined "narrative chains" as "partially ordered set[s] of narrative events that share a common actor" and use pointwise mutual information (PMI) as a measure of event association strength. Subsequent work showed that using argument consistency as a criterion for event relatedness improved model predictiveness as measured by a narrative cloze test, i.e., predicting a missing event [21]. However, even when trained on the Gigaword Corpus, performance was surprisingly weak, with the average "ranked position" of over 1,050 under the best performing condition.

Subsequent work introduced skip grams to compensate for data sparseness, language modeling formalisms better suited to cloze prediction (e.g., bigram probability rather than PMI), and recall@n rather than average rank as an evaluation metric [22]. A separate approach applied multiple sequence alignment to event sequences then extracting and simplifying the graph formed by treating each row as a node and adding edges to pairs of nodes that contain events that were consecutive in some event sequence [23].

Improved narrative cloze performance results were obtained by stricter constraints on multi-argument consistency [24], topic-specific training sets [25], and alternative language models, e.g., Hidden-Markov [26], Log-Bilinear [27], and Association Rule models [28]. However, no significant efforts appear to have been directed to the task of induction of legal narrative schemas or the use of such schemas in fact elicitation.

3. RIM: An Architecture for Narrative-Driven Case Elicitation

A system for narrative schema-based fact elicitation must perform two functions: acquiring narrative schemas from examples; and using those schemas to guide interactions with litigants. Figure 1 sets forth an architecture that performs these two functions.

The left side of Figure 1 details an off-line mechanism for inducing schemas from narrative corpora, which is the primary focus of this paper. The right side of Figure 1 second depicts a real time component that uses these schemas to distinguish relevant from irrelevant utterances and to identify facts that could distinguish among legal schemas if confirmed or disconfirmed. Specifically, each litigant's utterance is converted into a sequence of events to be added to the event sequence derived from prior utterances. The combined sequence is then com-

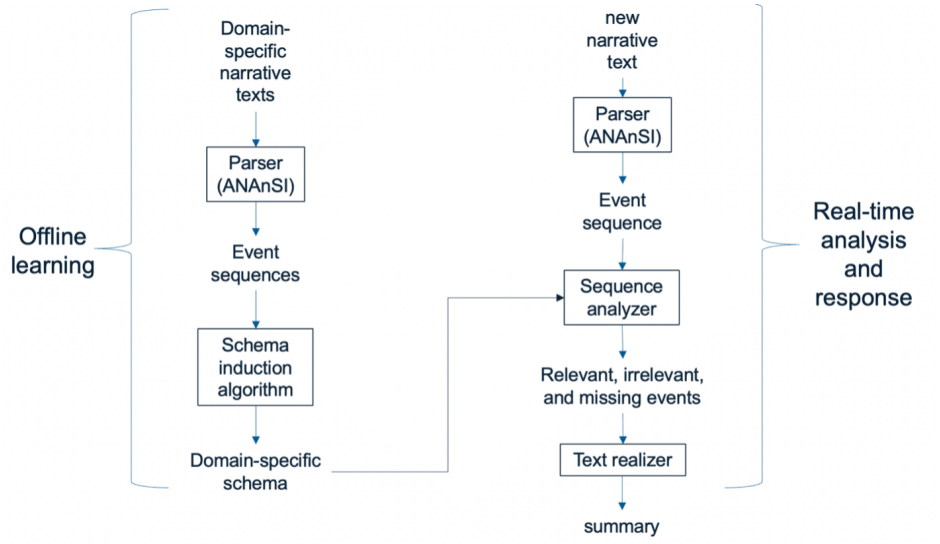

Figure 1: The RIM ("Relevant," "Irrelevant," and "Missing") architecture for narrative case elicitation.

pared to one or more narrative schemas. This comparison permits relevant events (those that match) to be distinguished from irrelevant events (unmatched events) and can suggest missing events (unmatched schema events) that should be inquired about. The Text Realizer generates questions to determine whether missing events can be confirmed or disconfirmed. Additional events elicited in this manner can distinguish among partially matching narratives or refine the match to the most similar narrative.

The real-time processing depicted on the right side of Figure 1 depends on the existence of a narrative schema for each area of law for which facts are to be elicited.[1] The process of induction of schemata from event sequences, depicted on the left side of Figure 1, is detailed in the next section. However, both the offline and real-time aspects of depend on conversion of raw text into event sequences, as shown as the second and third steps on both sides of Figure 1.

We term this architecture *RIM*, short for "Relevant, Irrelevant, and Missing," since the key functionality of the system is identifying these three categories of events.

4. Text to Event Sequence Conversion

The first step in converting text to event sequences is to parse each sentence into individual events and, for each event, identify the entities that fill the semantic roles of that event. The next step is analyzing the relationships among events by resolving coreferences and determining the discourse relations among the events. Many alternative approaches could be used to perform these two steps; we use ANAnSI (Advanced Narrative Analytics System Infrastructure) [30], a system that integrates the output of the Stanford Core NLP [31] constituency parser and cTakes [32] into a temporal, causal, and intentional graph represented in Neo4j [33] (see Figure 2).

4.1. Graph Linearization

The resulting graph representation for a collection of one or more sentences is then linearized into an event sequence with arguments and semantic roles standardized in the manner proposed in [24] to three alternative roles: agent, patient, and other complement. For example, in the event sequence shown in Figure 3, the pronouns "I" and "me" are normalized to "I".

4.2. Lemma Normalization

As discussed below in Section 5, corpora of legal narratives are, in general, many orders of magnitude smaller than the corpora used in previous narrative schema elicitation research, such as the Gigaword corpus [34]. Such small corpora produce sparse transition matrices with little predictive value, e.g., most event pairs in a new (or held out) event sequence will have never been seen before, meaning that there is no frequency data on which to base cloze predictions.

Several normalizations were therefore applied to reduce vocabulary size to improve matching. The most

[1] This process is described [29]

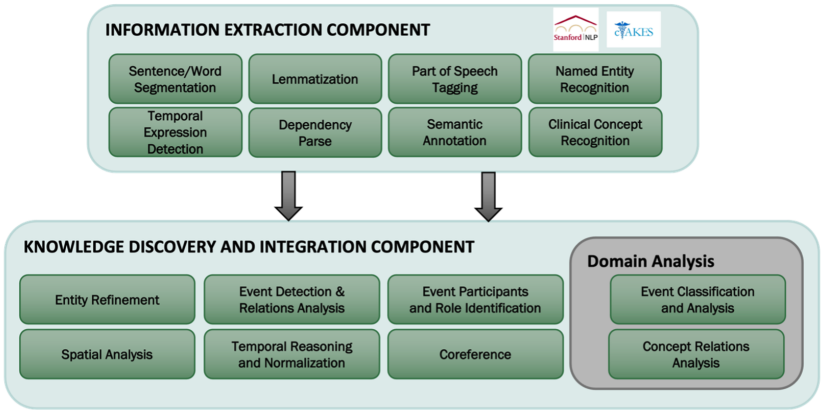

Figure 2: ANAnSI's NLP pipeline.

```
1427::be engineer(I, -, -),
1416::employment(I, -, -),
1424::place(Respondent, I, ['my employment', 'a leave of absence']),
1423::require(Respondent, I, -),
1418::pass(I, a medical exam, -),
1421::return(-, -, ['work']),
1408::file(I, a Charge of Discrimination, -),
1405::present(Respondent, I, ['a severance agreement']),
1413::unlawful(I, -, -),
1414::be unlawful(included terms, -, -),
1403::discharge(-, I, -),
```

Figure 3: A linearization of ANAnSI's temporal, causal, and intentional graph.

```
'intimidate': ['harass', 'intimidate', 'threaten'],
'keep': ['keep', 'leave'],
'look': ['feel', 'look'],
'make': ['give', 'make', 'put', 'take'],
'meet': ['meet', 'meeting'],
'regard': ['regard', 'respect'],
'send': ['email', 'send'],
'shift': ['change', 'shift'],
'speak': ['speak', 'talk'],
'suspend': ['revoke', 'suspend'],
'tell': ['ask', 'hear', 'know', 'let', 'tell'],
```

Figure 4: Lemma normalization by clustering event types in semantic embedding space.

important and general of these was lemma normalization, which consists of clustering events in semantic embedding space[2] and replacing each event with the most central member of the cluster in which it occurs. For example, Figure 4 show the results of complete-linkage hierarchical clustering of events in the EEOC (Equal Employment Opportunity Commission)[3] corpus (described below) with a minimum cosine threshold of 0.75.[4] For example, both "harass" and "threaten" are replaced by "intimidate," and 'ask," "hear," "know," and "let" are all replaced by "tell.' '

A second normalization that was motivated by the EEOC domain but useful in other domains was to replace each occurrence of a form of "to be" that has as an argument the name of an occupation with the event "be OCCUPATION."[5]

Lemma normalization shrinks the vocabulary size of the narrative, increasing transition matrix density and therefore increasing the likelihood that event cooccurrences will have been observed in the training corpus. This reduction in vocabulary size comes at the cost of reducing the specificity of the event representation.

5. Corpora

A key challenge for narrative schema-based case elicitations is the difficulty of obtaining significant numbers of narrative texts representative of narratives produced by litigants. In general, such texts contain sensitive personal information that precludes sharing in the form of public corpora. Documents filed in legal or administrative bodies are typically public, so statements of facts in petitions, complaints, and other filings can be a source of narrative texts. However, counsel for litigants often draft the statements in facts of court filings, so the text of such statements seldom contains language used by litigants

[2] We used the spaCy large English model, https://spacy.io/models/en.
[3] See https://www.eeoc.gov/.
[4] This threshold is an ad hoc setting, intended to be low enough to group synonymous terms without merging terms with obviously different meanings.
[5] We used the list of 1,156 occupations, from "accountant" to "zoologist" set forth in https://github.com/johnlsheridan/occupations/blob/master/occupations.csv. We used the occupational normalization in all experiments below.

themselves except in the case of self-represented (*pro se*) litigants, i.e., those who have no attorney to draft their statements in fact. The ideal corpus would consist of statements of fact in pro se litigants' filings, but such filings are difficult to obtain in bulk.

In this research, we obtained one small corpus of texts by pro se litigants together with five other data sets intended to reflect various characteristics of fact statements:

1. EEOC complaints. The complaints were transcribed from handwritten texts in the field titled "The facts supporting the plaintiff's claim of discrimination," in thirty employment discrimination complaints filed in the Northern District of Illinois in 2016. These texts are representative of litigant-generated narrative texts.
2. Multi-LexSum Summaries of Civil Cases. These 364 summaries of civil rights lawsuits were created for training and evaluating legal case summarization [35]. The Multi-LexSum text were included to typify procedural histories, a type of narrative required for appeals that court personnel have identified as being challenging for pro se appellants.
3. WIPO cases. The "background facts" of 6,000 decisions by World Intellectual Property Organization in domain name disputes. These fact statements were drafted by the panel deciding the case and are therefore not representative of pro se text. However, the similarity among these fact statements suggests that they could be a benchmark for narrative induction.
4. Board of Veteran Affairs decisions. The "Introduction" section of 1,680 Board of Veterans Appeals (BVA) cases. As with the WIPO cases, these texts are drafted by the judge writing the opinion and are therefore not representative of pro se text but potentially useful as a benchmark for narrative induction.
5. SPOT-HO online housing questions. Two hundred sixty three questions posed to the Suffolk University Law School's Legal Innovation and Technology (LIT) Lab issue spotting service [36].
6. SPOT-WO online employment questions. Two hundred ninety five employment questions posed to the SPOT site.

The size, type, and authors of each of the corpora are summarized in Table 1.

6. Experimental Evaluation

The RIM architecture, described above in Section 3, is based on the capability of a model trained on examples of legal narratives to guide interactions with a litigant based on distinguishing relevant from irrelevant facts as they are presented and predicting missing facts that would contribute to a coherent story. The relative effectiveness of narrative models for each of these activities can be estimated using narrative cloze tests, which estimate the ability of models to predict a missing (typically, the next) event in a sequence.

We explored two types of predictive models: language models; and event sequence alignment models.

6.1. Language Models

For each of the 6 data sets described above in Section 5 we converted each narrative text into a linearized event sequence, with events lemmatized by clustering in semantic vector space with a similarity threshold of 0.75. We calculated the recall@n in 10-fold cross validation. For these experiments, we relaxed the constraint that cooccurring events share common arguments to reduce the effects of data sparsity.

Several aspects of the results, shown in Table 2, suggest that data sparsity in narrative corpora of the magnitude of those evaluated in this experiment present a significant impediment to their use in the RIM framework. First, little improvement was observed between unigram and trigram models, suggesting that there are too few multi-event sequences for effective training. Moreover, there was only a modest improvement from recall@1 to recall@10, suggesting that many transitions in test data were never observed in the training data. Thus, data sparsity appears to remain a significant issue even after reducing the vocabulary size through semantic clustering.

6.2. Sequence Alignment Models

An alternative approach to narrative schema induction that may be more appropriate for domains with very sparse training data is based on *event sequence alignment*. In this approach, which is inspired by techniques of molecular biology, event sequences are aligned to find the most common subsequences, which can then be used as components of narrative schemas. Figure 5 shows the local alignment, that is, the alignment maximizing the longest common subsequence (LCS)[37], between two event sequences from the BVA corpus. Normalized LCS (NLCS) as shown in Formula 1 is a metric useful for comparing and grouping similar event sequences, even if they differ in lengths.

$$1.0 - (|LCS(s_1, s_2)|/mean(|s_1|, |s_2|))) \quad (1)$$

Intuitively, a cluster of event sequences sharing common subsequences may have a family resemblance [38]

Corpus	Size	Text Type	Author Type
EEOC	30	complaints	pro se litigant
SPOT-WO	295	legal advice requests	lay public
SPOT-HO	263	legal advice requests	lay public
Multi-Lexum	364	procedural history	federal judge
BVA	1,680	background facts	administrative law judge
WIPO	6,000	background facts	administrative law judge

Table 1
The size, type, and authors of each corpus of narrative texts.

Corpus	Model	Recall@1	Recall@5	Recall@10	Cloze score
EEOC	random	0.0084	0.0311	0.0467	109.7
	unigram LM	0.2877	0.2877	0.3199	30.7
	bigram LM	0.2877	0.2877	0.3199	30.7
	trigram LM	0.2879	0.2967	0.3243	30.9
Multi-LexSum	random	0.0019	0.0056	0.0147	376.3
	unigram LM	0.0669	0.0676	0.0713	208.3
	bigram LM	0.0669	0.0676	0.0713	208.3
	trigram LM	0.0838	0.0976	0.1023	202.0
SPOT-HO	random	0.0010	0.0044	0.0060	1081.9
	unigram LM	0.1343	0.1960	0.1986	431.2
	bigram LM	0.1343	0.1960	0.1986	431.2
	trigram LM	0.1305	0.1995	0.2040	421.4
SPOT-WO	random	0.0005	0.0023	0.0049	970.7
	unigram LM	0.1193	0.1214	0.1250	459.6
	bigram LM	0.1193	0.1214	0.1250	459.6
	trigram LM	0.1183	0.1232	0.1276	455.7
BVA	random	0.0016	0.0068	0.0130	389.3
	unigram LM	0.0000	0.0777	0.1541	245.0
	bigram LM	0.0000	0.0777	0.1541	245.0
	trigram LM	0.0000	0.1088	0.1871	234.7
WIPO	random	0.0011	0.0033	0.0059	974.2
	unigram LM	0.0001	0.1100	0.1298	528.4
	bigram LM	0.0001	0.1100	0.1298	528.4
	trigram LM	0.0001	0.1225	0.1472	513.1

Table 2
Comparison of event prediction performance using 1–3–gram language models in six legal narrative datasets.

that makes them useful for recognizing new event subsequences, e.g., that might share different subsequences with different cluster members. Consistent with this intuition, we perform the following steps to convert each corpus to a model consisting of a set of schema:

1. **Cluster**. Identify groups of sequences sharing common subsequences.
2. **Merge**. Create individual models from each group of sequences.
3. **Match**. Use each model to distinguish relevant, irrelevant, and missing events from new sequences.

Each of these steps is described in turn below.

6.2.1. Alignment-Based Sequence Clustering

Multiple sequence alignment is quite computationally expensive for collections of sequences in the size range of the 6 corpora in our experiments (30-6,000), so we use a heuristic approach derived from the center star alignment algorithm of [39].

1. Convert narratives to event sequences, as per Section 4.
2. Perform total-linkage agglomerative hierarchical clustering with distance metric NLCS and distance threshold t. The resulting clusters comprise event sequences that share a significant proportion of event subsequences.

As shown in Figure 6, achieving a mean cluster size of 2.0 requires a very high distance threshold, ranging from al-

	Event Sequence 1		Event Sequence 2
	operate	|	operate
	duty	|	duty
	bring	|	bring
	Appeal	|	Appeal
	decision	|	decision
−	problem	|	
	plead	|	plead
		| +	Hearing
	undersign	|	undersign
	Hearing	|	Hearing
	affiliate	|	affiliate
	remand	|	remand
		| +	problem
		| +	expertise
		| +	return
		| +	revisit

Figure 5: Local alignment (alignment maximizing the longest common subsequence) between two event sequences from the BVA corpus.

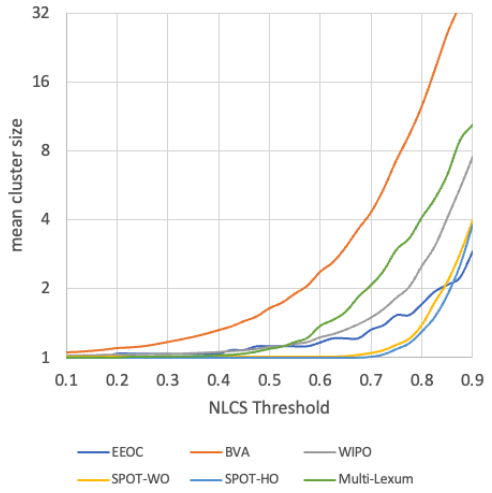

Figure 6: Mean number of cluster members as a function of t, the NLCS total-linkage distance threshold.

most 0.6 for the BVA corpus to over 0.8 for the SPOT-HO, SPOT-WO, and EEOC corpora. This is another indication that all these datasets are sparse and heterogeneous.

6.2.2. Merging Event Sequences

For each cluster, C, of similar event sequences, we perform the following steps to merge the sequences into a schema:

1. Identify the medoid, i.e., the sequence having the highest mean similarity to the other members of the cluster, breaking ties in favor of shorter sequences.
2. Convert the medoid sequence into a directed acyclic graph (DAG) in which each node is an event and each cooccurring pairs of events is connected by an edge from the earlier to the later event.
3. For each non-medoid sequence, S, in the cluster, align and merge S with the DAG by combining each node n of S with corresponding node in the best matching path in C. If n is unmatched, it is added as a new node in the DAG.

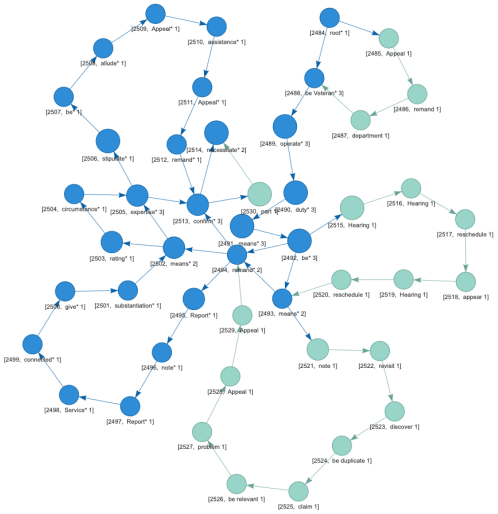

Figure 7: A DAG composed of 3 BVA event sequences

An example of the result of merging three BVA event sequences is shown in Figure 7. In the detail shown in Figure 8, events in the medoid are in blue, and each node is labeled with the number of sequences it occurs in, e.g., all three sequences contain a subsequence that starts with "be veteran," "operate," and "duty," but in only one sequence were these events preceded by "appeal," "remand," and "department."

The event sequence DAG created by this process is a model that makes recurring event subsequences explicit . Intuitively, a new sequence might strongly match multiple subsequences of the DAG even if there were other subsequences that were unmatched.

6.2.3. DAG Matching

In our initial procedure for matching a new sequence S with a DAG, we identified the alignment between S and each unique path in the DAG to find the path that maximizes the matched portion of S. We hypothesized that we would typically obtain a better match from the DAG than from any one of the individual events sequences merged into that DAG.

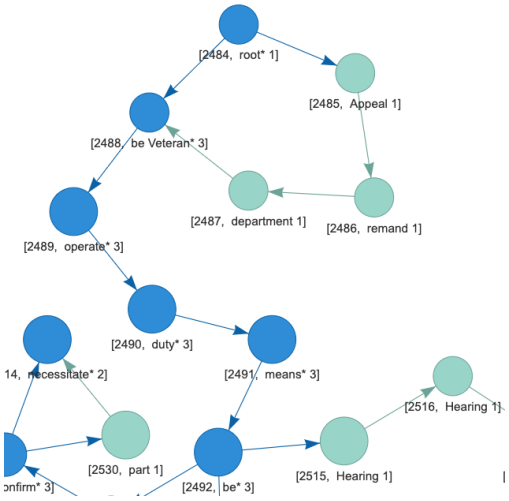

Figure 8: Detail of the 3 BVA event sequence DAG

6.2.4. Evaluation

We performed a preliminary evaluation of sequence alignment models using each of the six corpora. For each corpus, we clustered the event sequences as described above. Each event sequence S of each cluster C was held out for testing and the remaining members of the cluster, $(C - S)$, merged to form DAG_{C-S}.

Cloze test instances were created by replacing in turn each event $e \in S$ with a token, t, guaranteed not to be in the corpus (e.g., the token "$missing$"), aligning the resulting test sequence, $S[e \rightarrow t]$, with a model sequence, M, (models are described below), determining whether e was identified as missing in the LCS and, if so, counting how many other events in the model sequence were also identified as missing in the LCS. The proportion of events $e \in S$ identified as missing from $S[e \rightarrow t]$ under this procedure is the *recall* for S, and the *precision* is e's proportion of all missing events for all choices of e (i.e., for each e identified as missing, how many other events were also missing, meaning that they occur in M but not $S[e \rightarrow t]$).

We performed the cloze test procedure for each corpus under two conditions. In the test condition, the model sequence, M, consisted of the best matching path in DAG_{C-S}. In the control condition, the model sequence consisted of the sequence in $C - S$ that best matched S. The results of this experiment are shown in Table 3.

The first row is the distance threshold t required to ensure that the mean size of the clusters produced by hierarchical agglomerative clustering is at least 2. The second row, "proportion clustered," consists of the proportion of event sequences in each corpus that were included in some cluster containing at least 3 members (the minimum number to compare the DAG matching with matching to individual cluster members). The third row, "compression," represents the number of nodes in the DAG divided by the number of events in the event sequences composing that DAG, e.g., the proportion of overlap among the event sequences. The remaining rows show the precision, recall, and f-measure under the control condition (the model consisted simply of whatever case event sequence had the highest LCS) and the test condition (the model was the best DAG path with the highest LCS).

6.2.5. Results

The experimental evaluation showed somewhat surprisingly that slightly higher f-measure was obtained when the model was simply the event sequence in a cluster with the highest LCS with the test sequence rather than the best path in the DAG. Both sequence alignment methods produced better results than the language-model approach, but even the highest f-measures (0.132 and 0.125 for EEOC and WIPO, respectively) may not be sufficient for practical narrative elicitation applications. Note that narrative sequences that had very low similarity to any other sequence were not included in the evaluation (i.e., they were excluded from the "proportion clustered" in Table 3).

7. Summary and Discussion

This paper has proposed a new computational architecture for narrative-guided case elicitation, assembled six new legal narrative corpora, and evaluated two different approaches to creating legal narrative schemas, the first using language models, and the second using event sequence alignment. The cloze prediction accuracy observed in the language model approach was similar to previous narrative schema induction experiments, but we are skeptical that this performance is adequate for the needs of the RIM model. For example, recall@5 in the BVA corpus using a trigram model was only about 0.101, meaning that there would be only about a 10% chance that a missing event would be among the 5 most probable events as predicted by the model.

The evaluation of sequence alignment approach suggests that this approach may be more appropriate for sparse legal corpora such as the six corpora explored in this research. The initial observation that the best cloze performance can be obtained by simply using the most similar individual prior event sequence as a model, at least with these six corpora, is surprising but is not inconsistent with the observation that exemplar-based reasoning often works better with sparse datasets than more aggressive generalizers.

	EEOC	BVA	WIPO	SPOT-WO	SPOT-HO	Multi-Lexum
threshold	0.80	0.70	0.75	0.85	0.85	0.80
proportion clustered	0.38	0.83	0.42	0.44	0.31	0.86
compression	0.78	0.45	0.71	0.87	0.89	0.73
test recall	0.483	0.681	0.495	0.331	0.313	0.476
test precision	0.096	0.135	0.105	0.042	0.018	0.055
test f-measure	0.151	0.189	0.174	0.079	0.035	0.099
control recall	0.463	0.592	0.46	0.343	0.315	0.47
control precision	0.075	0.059	0.071	0.024	0.017	0.051
control f-measure	**0.132**	**0.103**	**0.125**	**0.046**	**0.033**	**0.094**

Table 3
Precision, recall, and f-measure in cloze tests applied to all clusters for which $|C| \geq 3$.

The work described in this paper is only an initial step in the research program of narrative-guided fact elicitation for self-represented litigants. Acquisition by the research community of larger datasets of legal narratives, particularly those produced by self-represented litigants, is a vital next step for progress in this important problem.

Acknowledgments

The authors express their gratitude to Charlotte Alexander for permitting us to experiment with her collection of EEOC complaints and to Charles Horowitz for assistance with ANAnSI. The MITRE Corporation is a not-for-profit company, chartered in the public interest. This document is approved for Public Release; Distribution Unlimited. Case Number 23-1184.

References

[1] C. E. Cerniglia, The civil self-representation crisis: The need for more data and less complacency, Georgetown Journal on Poverty Law and Policy 27 (2020) 355–388.

[2] S. Moore, A. Nwebury, Legal aid in crisis: Assessing the impact of reform, 1st ed., Bristol University Press, 2017. https://doi.org/10.2307/j.ctt1t8988q.

[3] G. C. Grimwood, Litigants in person: the rise of the self-represented litigant in civil and family cases in england and wales, https://commonslibrary.parliament.uk/research-briefings/sn07113/, 2016. House of Commons Library.

[4] A. Biard, J. Hoevenaars, X. Kramer, E. Themeli, Introduction: The future of access to justice—beyond science fiction, in: X. Kramer, A. Biard, J. Hoevenaars, E. Themeli (Eds.), New Pathways to Civil Justice in Europe, Springer, Cham, 2021, pp. 1–20.

[5] J. Macfarlane, The national self-represented litigants project: Identifying and meeting the needs of self-represented litigants: Final report, CanLIIDocs (2013) 493.

[6] K. Scarrow, B. Fragomeni, J. Macfarlane, Tracking the trends of the self-represented litigant phenomenon: Data from the national self-represented litigants project, 2018/2019, 2020.

[7] K. Joyce, No money, no lawyer, no justice, The New Republic (2002) 38–45.

[8] K. M. Kroeper, V. D. Quintanilla, M. Frisby, N. Y. A. G. Applegate, S. J. Sherman, M. C. Murphy, Underestimating the unrepresented: Cognitive biases disadvantage pro se litigants in family law cases, Psychology, Public Policy, and Law 26 (2020) 198–212.

[9] D. Himonas, T. Hubbard, Democratizing the rule of law, Stanford Journal of Civil Rights & Civil Liberties 16 (2020) 261–282.

[10] A. Schmitz, J. Zeleznikow, Intelligent legal tech to empower self-represented litigants, Columbia Science and Technology Law Review (2022) 142–190.

[11] M. Lauritsen, Q. Steenhuis, Substantive legal software quality: A gathering storm?, in: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, ICAIL '19, ACM, New York, NY, USA, 2019, pp. 52–62.

[12] J. Morgan, A. Paiement, M. Seisenberger, J. Williams, A. Wyner, A chatbot framework for the children's legal centre, in: M. Palmirani (Ed.), Legal Knowledge and Information Systems - JURIX 2018, volume 313 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2018, pp. 205–209.

[13] M. J. Sergot, F. Sadri, R. A. Kowalski, F. Kriwaczek, P. Hammond, H. T. Cory, The British Nationality Act as a logic program, Commun. ACM 29 (1986) 370–386. URL: http://doi.acm.org/10.1145/5689.5920. doi:10.1145/5689.5920.

[14] C. D. Phillips, Reconstructing reality in the courtroom: Justice and judgment in american culture, American Political Science Review 77 (1983) 275–276. doi:10.2307/1956108.

[15] N. Pennington, R. Hastie, A cognitive theory of juror decision making: The story model, Cardozo L. Rev. 13 (1991) 519.

[16] N. Pennington, R.Hastie, Explaining the evidence: Tests of the story model for juror decision making, Journal of Personality and Social Psychology 62 (1992) 189–206. https://doi.org/10.1037/0022-3514.62.2.18.

[17] F. D. Donato, Fact-finding in contexts: framing clients' agentivity within judicial and administrative procedures, file:///Users/lbranting/Downloads/DIDONATO_position_paper_Rotterdam_22.9.15.pdf, last accessed February 5, 2023, 2015.

[18] M. Livingston, A Sad Story in Not a Legal Defense: Defining Legal Issues, Master's thesis, University of Colorado at Boulder Dept. of Communication, 2014.

[19] R. C. Schank, R. Abelson, Scripts, Plans, Goals, and Understanding, Hillsdale, NJ: Earlbaum Assoc, 1977.

[20] N. Chambers, D. Jurafsky, Unsupervised learning of narrative event chains, in: Proceedings of ACL-2007, Association for Computational Linguistics, Columbus, Ohio, 2008, pp. 789–797.

[21] N. Chambers, D. Jurafsky, Unsupervised learning of narrative schemas and their participants, in: Proceedings ACL-2009, Association for Computational Linguistics, USA, 2009, p. 602–610.

[22] B. Jans, S. Bethard, I. Vulic, M.-F. Moens, Skip n-grams and ranking functions for predicting script events, in: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012), ACL; East Stroudsburg, PA, 2012, pp. 336–344.

[23] M. Regneri, A. Koller, M. Pinkal, Learning script knowledge with web experiments, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010, pp. 979–988.

[24] K. Pichotta, R. Mooney, Statistical script learning with multi-argument events, in: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014, pp. 220–229.

[25] E. Rahimtoroghi, E. Hernandez, M. A. Walker, Learning fine-grained knowledge about contingent relations between everyday events, arXiv preprint arXiv:1708.09450 (2017).

[26] J. W. Orr, P. Tadepalli, J. R. Doppa, X. Z. Fern, T. G. Dietterich, Learning scripts as hidden markov models, CoRR abs/1809.03680 (2018).

[27] R. Rudinger, P. Rastogi, F. Ferraro, B. Van Durme, Script induction as language modeling, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1681–1686.

[28] A. Belyy, B. Van Durme, Script induction as association rule mining, in: Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events, Association for Computational Linguistics, Online, 2020, pp. 55–62.

[29] L. K. Branting, S. McLeod, Narrative-driven case elicitation, in: Proceedings of the ICAIL 2023 Workshop on AI for Access to Justice, Braga, Portugal, 19 June 2023.

[30] K. Megerdoomian, K. Branting, C. Horowitz, A. Marsh, S. Petersen, E. Scott, Automated narrative extraction from administrative records., in: AIAS@ ICAIL, 2019, pp. 38–48.

[31] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, D. McClosky, The stanford corenlp natural language processing toolkit., in: ACL (System Demonstrations), The Association for Computer Linguistics, 2014, pp. 55–60.

[32] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, C. G. Chute, Mayo clinical text analysis and knowledge extraction system (ctakes), JAMIA 17 (2010) 507–513.

[33] Neo4j, Neo4j -graph data platform, 2023. http://neo4j.com/, last accessed 7 March 2023.

[34] C. Napoles, M. R. Gormley, B. Van Durme, Annotated gigaword, in: Proceedings of the joint workshop on automatic knowledge base construction and web-scale knowledge extraction (AKBC-WEKEX), 2012, pp. 95–100.

[35] Z. Shen, K. Lo, L. J. Yu, N. Dahlberg, M. Schlanger, D. Downey, Multi-lexsum: Real-world summaries of civil rights lawsuits at multiple granularities, ArXiv abs/2206.10883 (2022).

[36] SPOT, The legal innovation & technology lab's spot api, https://spot.suffolklitlab.org, last accessed March 9, 2023, 2023.

[37] D. Gusfield, Algorithms on Strings, Trees, and Sequences, Cambridge University Press, 1999.

[38] L. Wittgenstein, Philosophical investigations. Philosophische Untersuchungen, Macmillan, 1953.

[39] Q. Zou, M.-z. GUO, X.-k. WANG, T.-t. ZHANG, An algorithm for dna multiple sequence alignment based on center star method and keyword tree, ACTA ELECTONICA SINICA 37 (2009) 1746.