APPENDIX 1011D

aclanthology.org/2022.findings-emnlp.447/

Sign in

ACL Anthology News FAQ Corrections Submissions Github

Search...

Validity Assessment of Legal Will Statements as Natural Language Inference

Alice Kwak, Jacob Israelsen, Clayton Morrison, Derek Bambauer, Mihai Surdeanu

Abstract

This work introduces a natural language inference (NLI) dataset that focuses on the validity of statements in legal wills. This dataset is unique because: (a) each entailment decision requires three inputs: the statement from the will, the law, and the conditions that hold at the time of the testator's death; and (b) the included texts are longer than the ones in current NLI datasets. We trained eight neural NLI models in this dataset. All the models achieve more than 80% macro F1 and accuracy, which indicates that neural approaches can handle this task reasonably well. However, group accuracy, a stricter evaluation measure that is calculated with a group of positive and negative examples generated from the same statement as a unit, is in mid 80s at best, which suggests that the models' understanding of the task remains superficial. Further ablative analyses and explanation experiments indicate that all three text segments are used for prediction, but some decisions rely on semantically irrelevant tokens. This indicates that overfitting on these longer texts likely happens, and that additional research is required for this task to be solved.

PDF

Cite

Search

Video

Fix data

Anthology ID: 2022.findings-emnlp.447
Volume: Findings of the Association for Computational Linguistics: EMNLP 2022
Month: December
Year: 2022
Address: Abu Dhabi, United Arab Emirates
Editors: Yoav Goldberg, Zornitsa Kozareva, Yue Zhang
Venue: Findings
SIG: –
Publisher: Association for Computational Linguistics
Note: –
Pages: 6047–6056

Pages: 6047–6056
Language: –
URL: https://aclanthology.org/2022.findings-emnlp.447/
DOI: 10.18653/v1/2022.findings-emnlp.447
Bibkey: kwak-etal-2022-validity
Cite (ACL): Alice Kwak, Jacob Israelsen, Clayton Morrison, Derek Bambauer, and Mihai Surdeanu. 2022. Validity Assessment of Legal Will Statements as Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6047–6056, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
Cite (Informal): Validity Assessment of Legal Will Statements as Natural Language Inference (Kwak et al., Findings 2022)
Copy Citation: BibTeX Markdown MODS XML Endnote More options...
PDF: https://aclanthology.org/2022.findings-emnlp.447.pdf
Video: https://aclanthology.org/2022.findings-emnlp.447.mp4

ACL materials are Copyright © 1963–2026 ACL; other materials are copyrighted by their respective copyright holders. Materials prior to 2016 here are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 International License. Permission is granted to make copies for the purposes of teaching and research. Materials published in or after 2016 are licensed on a Creative Commons Attribution 4.0 International License.

The ACL Anthology is managed and built by the ACL Anthology team of volunteers.

Site last built on 06 January 2026 at 16:21 UTC with commit 955213d.

ACL Anthology

- News
- FAQ
- Corrections
- Submissions
- Github

Search...

Validity Assessment of Legal Will Statements as Natural Language Inference

Alice Kwak, Jacob Israelsen, Clayton Morrison, Derek Bambauer, Mihai Surdeanu

Correct Metadata for

×

Important: The Anthology treat PDFs as authoritative. Please use this form only to correct data that is out of line with the PDF. See our corrections guidelines if you need to change the PDF.

Title Adjust the title. Retain tags such as <fixed-case>.

Authors Adjust author names and order to match the PDF.

Add Author

Abstract Correct abstract if needed. Retain XML formatting tags such as <tex-math>.

Verification against PDF Ensure that the new title/authors match the snapshot below. (If there is no snapshot or it is too small, consult the PDF.)

Authors concatenated from the text boxes above:

ALL author names match the snapshot above—including middle initials, hyphens, and accents.

Submit

Abstract

This work introduces a natural language inference (NLI) dataset that focuses on the validity of statements in legal wills. This dataset is unique because: (a) each entailment decision requires three inputs: the statement from the will, the law, and the conditions that hold at the time of the testator's death; and (b) the included texts are longer than the ones in current NLI datasets. We trained eight neural NLI models in this dataset. All the models achieve more than 80% macro F1 and accuracy, which indicates that neural approaches can handle this task reasonably well. However, group accuracy, a stricter evaluation measure that is calculated with a group of positive and negative examples generated from the same statement as a unit, is in mid 80s at best, which suggests that the models' understanding of the task remains superficial. Further ablative analyses and explanation experiments indicate that all three text segments are used for prediction, but some decisions rely on semantically irrelevant tokens. This indicates that overfitting on these longer texts likely happens, and that additional research is required for this task to be solved.

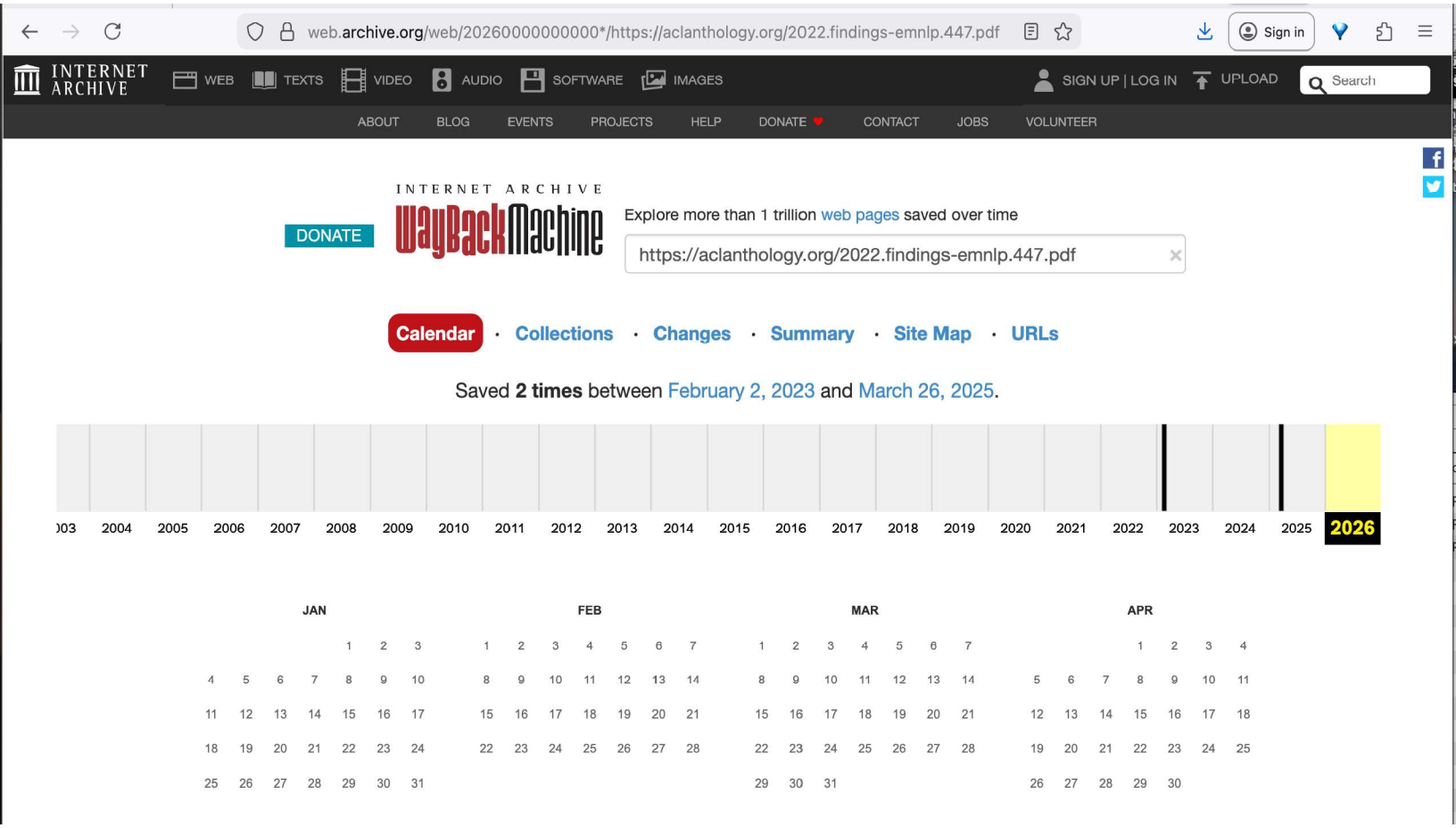
Anthology ID:
2022.findings-emnlp.447
Volume:
Findings of the Association for Computational Linguistics: EMNLP 2022
Month:
December
Year:
2022
Address:

Abu Dhabi, United Arab Emirates
Editors:
Yoav Goldberg, Zornitsa Kozareva, Yue Zhang
Venue:
Findings
SIG:
Publisher:
Association for Computational Linguistics
Note:
Pages:
6047–6056
Language:
URL:
https://aclanthology.org/2022.findings-emnlp.447/
DOI:
10.18653/v1/2022.findings-emnlp.447
Bibkey:
kwak-etal-2022-validity
Cite (ACL):
Alice Kwak, Jacob Israelsen, Clayton Morrison, Derek Bambauer, and Mihai Surdeanu. 2022. Validity Assessment of Legal Will Statements as Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6047–6056, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
Cite (Informal):
Validity Assessment of Legal Will Statements as Natural Language Inference (Kwak et al., Findings 2022)
Copy Citation:
BibTeX Markdown MODS XML Endnote More options...
PDF:
https://aclanthology.org/2022.findings-emnlp.447.pdf
Video:
https://aclanthology.org/2022.findings-emnlp.447.mp4

ACL materials are Copyright © 1963–2026 ACL; other materials are copyrighted by their respective copyright holders. Materials prior to 2016 here are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 International License. Permission is granted to make copies for the purposes of teaching and research. Materials published in or after 2016 are licensed on a Creative Commons Attribution 4.0 International License.

The ACL Anthology is managed and built by the ACL Anthology team of volunteers.

Site last built on 06 January 2026 at 16:21 UTC with commit 955213d.

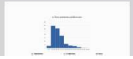

web.archive.org/web/20260000000000*/https://aclanthology.org/2022.findings-emnlp.447.pdf
Sign in
INTERNET ARCHIVE
WEB TEXTS VIDEO AUDIO SOFTWARE IMAGES
SIGN UP | LOG IN UPLOAD
Search
ABOUT BLOG EVENTS PROJECTS HELP DONATE CONTACT JOBS VOLUNTEER
DONATE
INTERNET ARCHIVE
WayBack Machine
Explore more than 1 trillion web pages saved over time
https://aclanthology.org/2022.findings-emnlp.447.pdf
Calendar · Collections · Changes · Summary · Site Map · URLs
Saved 2 times between February 2, 2023 and March 26, 2025.
003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 2026
JAN
1 2 3
4 5 6 7 8 9 10
11 12 13 14 15 16 17
18 19 20 21 22 23 24
25 26 27 28 29 30 31
FEB
1 2 3 4 5 6 7
8 9 10 11 12 13 14
15 16 17 18 19 20 21
22 23 24 25 26 27 28
MAR
1 2 3 4 5 6 7
8 9 10 11 12 13 14
15 16 17 18 19 20 21
22 23 24 25 26 27 28
29 30 31
APR
1 2 3 4
5 6 7 8 9 10 11
12 13 14 15 16 17 18
19 20 21 22 23 24 25
26 27 28 29 30

web.archive.org/web/20230202142404/https://aclanthology.org/2022.findings-emnlp.447.pdf
Sign in
INTERNET ARCHIVE Wayback Machine
https://aclanthology.org/2022.findings-emnlp.447.pdf
Go
2 captures
2 Feb 2023 - 26 Mar 2025
JAN FEB MAR
2022 02 2023 2024
About this capture
6076 (1 of 10)
Automatic Zoom









Validity Assessment of Legal Will Statements as Natural Language Inference

Alice Saebom Kwak, Jacob O. Israelsen, Clayton T. Morrison, Derek E. Bambauer, Mihai Surdeanu
The University of Arizona, Tucson, Arizona, USA
{alicekwak, jisraelsen, claytonm, derekbambauer, msurdeanu}@arizona.edu

Abstract

This work introduces a natural language inference (NLI) dataset that focuses on the validity of statements in legal wills. This dataset is unique because: (a) each entailment decision requires three inputs: the statement from the will, the law, and the conditions that hold at the time of the testator's death; and (b) the included texts are longer than the ones in current NLI datasets. We trained eight neural NLI models

NLI in general deals with determining whether a premise entails, contradicts, or is neutral to the hypothesis given. Our project made two important changes to this approach to fit legal documents. First, our dataset contains three input types: statements from wills (as hypotheses), laws (as premises), and *conditions*, which are circumstances at the time of probate (e.g., eligibility of the testator or witnesses). This adaptation is necessary

Validity Assessment of Legal Will Statements as Natural Language Inference

Alice Saebom Kwak, Jacob O. Israelsen, Clayton T. Morrison, Derek E. Bambauer, Mihai Surdeanu
The University of Arizona, Tucson, Arizona, USA
{alicekwak, jisraelsen, claytonm, derekbambauer, msurdeanu}@arizona.edu

Abstract

This work introduces a natural language inference (NLI) dataset that focuses on the validity of statements in legal wills. This dataset is unique because: (a) each entailment decision requires three inputs: the statement from the will, the law, and the conditions that hold at the time of the testator's death; and (b) the included texts are longer than the ones in current NLI datasets. We trained eight neural NLI models in this dataset. All the models achieve more than 80% macro F1 and accuracy, which indicates that neural approaches can handle this task reasonably well. However, group accuracy, a stricter evaluation measure that is calculated with a group of positive and negative examples generated from the same statement as a unit, is in mid 80s at best, which suggests that the models' understanding of the task remains superficial. Further ablative analyses and explanation experiments indicate that all three text segments are used for prediction, but some decisions rely on semantically irrelevant tokens. This indicates that overfitting on these longer texts likely happens, and that additional research is required for this task to be solved.

1 Introduction

Natural language inference (NLI) in the legal domain has not been widely investigated, despite its importance and potential. One such important NLI application is validity assessment of legal documents such as wills. Legal procedures for creating and executing wills are evolving rapidly. Processing a will via probate is a costly, time-consuming process that can be exacerbated by errors or by challenges to validity. These problems will likely increase as people increasingly employ electronic wills. Developing natural language techniques that can determine a will's validity at creation and execution can increase validity, conserve legal resources, and effectuate the author's intent. To this end, this work explores how NLI models can be employed to evaluate the validity of will statements.

NLI in general deals with determining whether a premise entails, contradicts, or is neutral to the hypothesis given. Our project made two important changes to this approach to fit legal documents. First, our dataset contains three input types: statements from wills (as hypotheses), laws (as premises), and *conditions*, which are circumstances at the time of probate (e.g., eligibility of the testator or witnesses). This adaptation is necessary as will statements' validity cannot be evaluated without considering both circumstances at the time of probate ("conditions") and relevant legal rules ("laws"). Also, we switched the labels from entailment, contradict, neutral to support, refute, unrelated to better represent the relation between will statements and laws.
The major contributions of this work are:

- We create an open-access annotated dataset with 1,014 data points, generated from 23 publicly available wills. In addition to the three-tuple setting unique to the legal domain, this dataset also contains texts considerably longer than in other open-domain NLI datasets, such as SNLI (Bowman et al., 2015). The average length of our texts is 269 tokens, while the average lengths of premises and hypotheses in SNLI are 8 and 14 tokens, respectively.

- We demonstrate that validity assessment of legal will statements can be handled reasonably well with state-of-the-art NLI models when trained on our dataset. However, low scores in group accuracy, which is a stricter evaluation measure calculated with a group of positive and negative examples generated from the same statement as a unit, indicate that work remains before NLI models fully understand legal language.

- We explain how the trained models utilize our dataset via ablation tests that remove various

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6076 - 6085
Exhibit 1026 Page 131
December 7-11, 2022 ©2022 Association for Computational Linguistics

pieces of information at prediction time (laws, conditions, or both), and through post-hoc explainability analyses using LIME (Ribeiro et al., 2016). Our analyses indicate that in most cases the NLI classifiers use meaningful phrases from all three pieces of text, indicating the models do capture useful information. However, in some situations, the models use features that are not intuitive for humans, indicating that the task is incomplete.

2 Related Work

Natural language inference (NLI), also known as Recognizing Textual Entailment (RTE), determines entailment relations between a pair of sentences: a premise and a hypothesis. Their relationship is either: a) entailment, if a premise entails a hypothesis; b) contradiction, if a premise contradicts a hypothesis; or c) neutral, if a premise neither entails nor contradicts a hypothesis. NLI has been a key framework in natural language processing since Dagan et al. (2006) proposed the RTE challenge.

Numerous datasets exist for general NLI tasks. For example, the Stanford Natural Language Inference (SNLI) Corpus (Bowman et al., 2015) dataset contains about 570,000 pairs of sentences (premises and hypotheses) generated from Flickr image captions. More recently, datasets for domain-specific NLI tasks are being introduced. SciNLI (Sadat and Caragea, 2022) is drawn from scientific texts. It contains 107,412 sentence pairs extracted from papers in natural language processing and computational linguistics.

There have been a few NLI-related resources for the legal domain, including IFlyLegal (Wang et al., 2019), StAtutory Reasoning Assessment (SARA) dataset (Holzenberger et al., 2020), Graph-based Causal Inference (GCI) framework (Liu et al., 2021), and ContractNLI (Koreeda and Manning, 2021). One IFlyLegal module was developed for natural language article inference; it provides relevant legal articles when asked legal questions. SARA contains rules extracted from the US Internal Revenue Code and natural language questions. GCI generates causal graphs based on fact descriptions. Though the framework was not specifically designed for law, Liu et al. (2021) demonstrated that it can be utilized for legal text analysis. ContractNLI is a dataset consisting of 607 legal contracts designed to handle document-level natural language inference.

Our work differs from these existing NLI works in two ways: (a) it contains three types of input information; and (b) it operates in a pragmatic middle space with respect to text length: larger than datasets with sentence-level texts, which are insufficient to capture legal details, but shorter than document-level inference datasets.

3 Dataset

Our dataset includes three types of entries (will statements, applicable law, and facts representing external state) rather than the usual two types. Validity often depends upon such external circumstances. Thus, a will statement can be legally valid under some facts, but not others. Hence, a will can change from valid to invalid or vice versa; wills must, at minimum, be evaluated when executed (drafted and signed) and when probated (where a court determines whether to implement its provisions). For example, if a will contains a statement appointing a specific person as executor, that person's eligibility must be verified. This requires information about both the law and the person's circumstances at probate time. For example, according to Tennessee Code section 40-20-115, any person who has been imprisoned cannot be an executor. Thus, to assess eligibility of a Tennessee executor in Tennessee, one must know whether they have been in prison at probate time.

3.1 Data Collection

Our data was collected from the U.S. Wills and Probates datasets in Ancestry, which contains documents in the public domain.[1] We chose 23 wills based on three criteria: 1) whether the wills were handwritten; 2) the wills' execution and probation dates; and 3) where the wills were executed and probated. We excluded handwritten wills due to the difficulty of OCR text recognition. Execution and probation dates can affect interpretation of wills; we excluded wills executed before 1970 and probated before 2000. Lastly, we only collected wills from Tennessee, because including wills from multiple states would require analyzing each state's laws governing wills. Tennessee had the greatest number of wills meeting our criteria.

All collected wills were anonymized. Personal information was replaced with special tokens denoting the type of information (such as [Person-

[1]Court documents are in the public domain in the US.

Exhibit 1026 Page 132

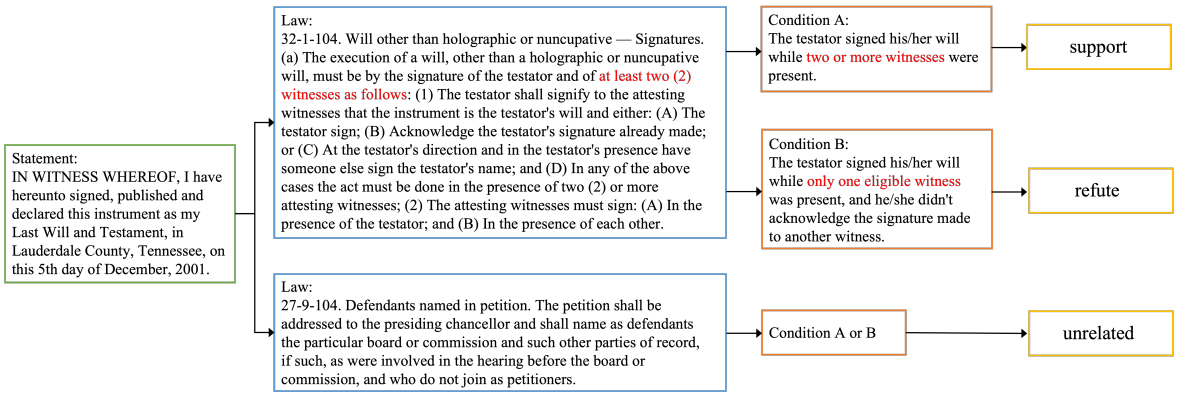


Figure 1: A demonstration of the annotation procedure. Given a law and condition, each statement is evaluated for validity. *Support*: it is supported by the law and condition provided; *Refute*: refuted by the law and condition. *Unrelated*: not relevant to the law. When a statement and a law are unrelated to each other, a condition is not required for classification. However, it is still assigned one of the conditions from either support or refute case. The goal is to prevent NLI models from depending on contextual difference (i.e., "texts without conditions are all unrelated cases") rather than features when making predictions.

n], [Address-n], [Number-n], where n identified each person or object in the will. (Suntwal et al. (2019) suggested this anonymization method). Anonymization was performed manually to prevent including personal information in the dataset.

3.2 Annotation

Each will statement was annotated as *support*, *refute*, or *unrelated* based on a given condition and law. (Our annotators set hypothetical conditions when necessary to categorize statements.)

Importantly, each statement was annotated multiple times. A statement was annotated once as support, once as refute, and thrice as unrelated.[2] This ensured that every statement was labeled with all three classes in the same ratio (support:refute:unrelated=1:1:3). Also, annotating statements three times more with "unrelated" than support and refute enabled including a greater range of laws in the dataset. Formally, our annotation procedure included the following steps:

(1) preprocessing: extracting texts from collected wills using OCR[3], and copying statements into the dataset;

(2) for each statement, identifying the Tennessee legal provision that supported, refuted, or was unrelated to the statement's validity;

(3) adding a condition specifying external circumstances relevant to validity;

(4) repeating these steps to generate five annotations per statement;

(5) anonymizing all statements by replacing personal information with special tokens.

Figure 1 demonstrates the annotation procedure with examples.

Two annotators participated in the task. One is a law student, and annotation was supervised by author Bambauer, a law professor. The other annotator does not have legal training, but ongoing discussions and reviews ensured the uniform quality. The annotators contributed equally. Annotators worked individually but shared annotation guidelines. After the annotation was complete, cross-annotation was conducted on 200 randomly selected items (100 items from each annotator's set; each annotator worked blinded on items drawn from the other annotator's set) to check inter-annotator agreement. We calculated Kappa agreement score based on the cross-annotation result, and the score is 0.91 (rounded to hundredth).

After the annotation was complete, the dataset was split into training, development, and test sets. There are 1014 data points in our dataset, split 50:25:25% (train: 504, development: 255, test: 255). When splitting, a group containing annotations from a single statement (one support, one refute, and three unrelated) was treated as a single

[2] A few exceptions occurred when:
a) it was impossible to create a case where the statement was refuted. In such cases, statements were not annotated as "refute."
b) a statement had more than one relevant legal rule. Such statements were annotated twice as support, twice as refute, and once as unrelated.

[3] The collected wills were scanned documents, so OCR was necessary to extract machine-readable text.

Exhibit 1026 Page 133

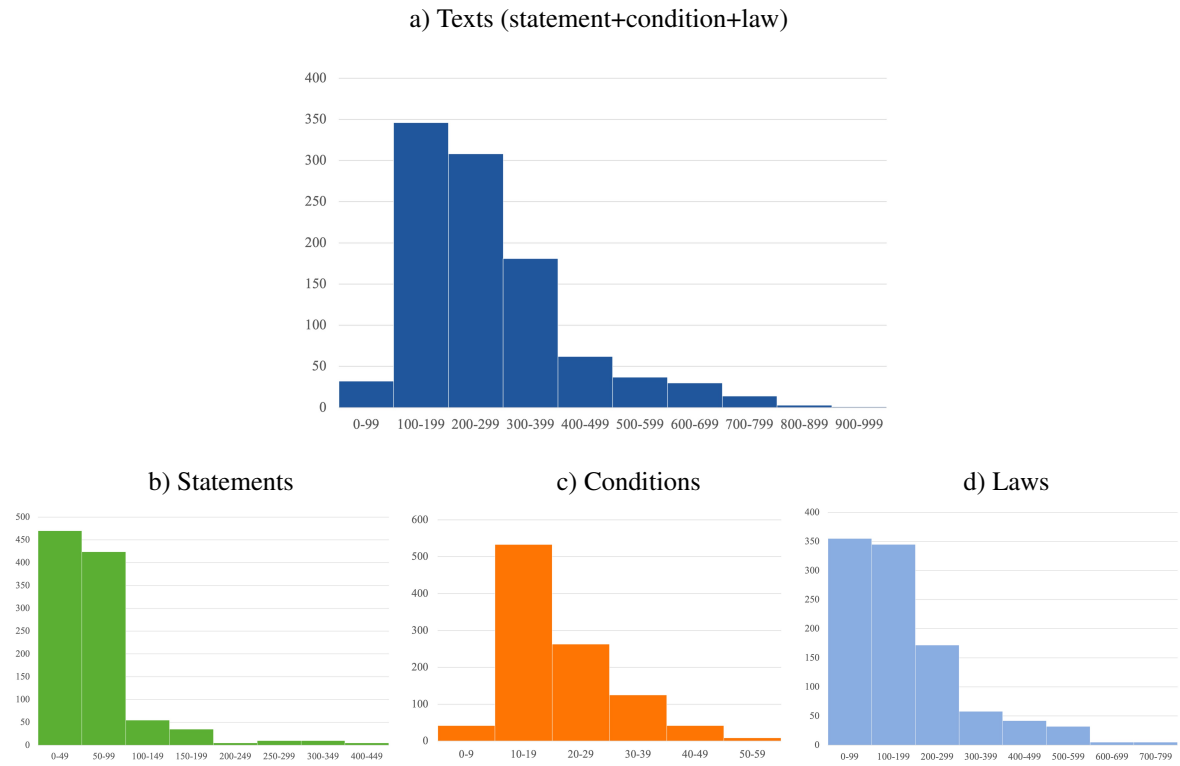


Figure 2: Histograms demonstrating the length distribution of a) full texts (i.e., statement+condition+law), b) statements, c) conditions, and d) laws in terms of token counts. Token counts are plotted on the X-axis; the Y-axis indicates the number of texts/statements/conditions/laws for each bin.

Datasets	**Texts**	**Tokens**
Train	504	135218
Development	255	69643
Test	255	68042
Total	**1014**	**272903**

Table 1: The number of texts and tokens contained in each dataset and in total

unit. Each annotator's work was equally represented in all three datasets. All 23 wills appear in more than one of the train, dev, and test partitions. However, given the nature of the will statements from the wills and how they were annotated, data leakage is unlikely. Will statements are independent and generally do not convey information about other statements from the same wills. Summary statistics are in Table 1 and Figure 2.

3.3 Characteristics of the Dataset

Our dataset has two characteristics distinguishing it from other datasets: a) texts are composed of three input types, and b) texts contain a large number of tokens.

First, our task requires a third input type: a condition, since validity often depends upon conditions.

Second, our texts tend to contain a large number of tokens. 44% of our texts contain 200 or more tokens. The average token number was 269.14; 79 texts contain more than 512 tokens (the threshold for most NLI models). Popular NLI datasets are shorter.

3.4 Implementation

We trained multiple NLI models with our dataset to assess their performance: four transformer models and four sentence-transformer models. The transformer models include bert-base-uncased (Devlin et al., 2018), distilbert-base-uncased (Sanh et al., 2019a), roberta-large-mnli (Liu et al., 2019), and longformer-base-4096 (Beltagy et al., 2020). Bert-base-uncased and dilstilbert-base-uncased were trained to set baselines on the task. In addition to the baseline models, we used roberta-large-mnli and longformer-base-4096. Among sentence-transformer models (Reimers and Gurevych, 2019), four pretrained models with top average performances (based on the Model Overview page[4]) were chosen: all-mpnet-base-v2[5], multi-qa-mpnet-base-

[4] https://www.sbert.net

[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Exhibit 1026 Page 134

dot-v1[6], all-distilroberta-v1 (Sanh et al., 2019b)[7], and all-MiniLM-L12-v2 (Wang et al., 2020)[8].

To distinguish between statements, laws, and conditions in the concatenated texts, we prefixed each with a special token: `[STATE]`, `[LAW]`, and `[COND]`.

Both transformer models and sentence-transformer models were trained on PyTorch 1.11.0 with Cuda 11.3. All transformer models were trained using the Trainer class provided by HuggingFace. Learning rates and training epochs were tuned on the development partition. Sentence-transformer models were trained with Sentence Transformer Fine-Tuning (SetFit) proposed by Wasserblat (2021). It was slightly adapted to fit the multi-class classification task, but its fundamentals remain intact. SetFit utilizes sentence pairs generated within the same class as training data. It fits the model with the data to minimize the Softmax Loss. Once the model is fitted, it is used to encode the training and development (or test, when it is a testing phase) datasets. The encoded data is used to fit the Logistic Regression classifier (with 'liblinear' solver) which, finally, makes predictions.

All models except for longformer-base-4096 were trained and evaluated on the truncated datasets due to the models' token number limitations. The models can only process texts with 512 or fewer tokens without truncation (all-MiniLM-L12-v2: 256 tokens; all-mpnet-base-v2: 384 tokens; other models, except longformer-base-4096: 512 tokens). When truncating datasets, the ratio of average token numbers among three input types (statements, laws, and conditions) was considered to prevent losing excessive information from a single type of input.

4 Results and Analysis

4.1 Evaluation Measures

We report results using standard accuracy, precision, recall, and F1 scores. To ensure all labels are well represented, precision, recall, and F1 scores were computed with Macro average. Additionally, as suggested by Elazar et al. (2021), we introduce a new measure called group accuracy (GA). GA is calculated with a group of positive and negative examples from the same statement (rather than each text) as a unit. If a group has one or more incorrect predictions, the group is incorrect. If a model correctly understands a will statement, it should perform equally well on all examples from the same statement.

4.2 Results from Trained Models

Table 2 shows the models' performances on our test partition. Overall, the models demonstrated good performances. Each achieved more than 80% in all metrics but group accuracy. Roberta-large-mnli showed the best performance. It achieved over 96% in all metrics but group accuracy (84.31%), suggesting it can handle the task reasonably well. However, the large difference between accuracy and group accuracy suggests that the models' understanding of the task is superficial.

Accuracy for the *unrelated* label was higher than for *support* and *refute* in all models except all-mpnet-base-v2 and multi-qa-mpnet-base-dot-v1. This higher accuracy for *unrelated* label may be partially attributable to the gap between accuracy and group accuracy, as it would inflate overall accuracy. However, since the gap between accuracy and group accuracy is not significantly smaller for all-mpnet-base-v2 and multi-qa-mpnet-base-dot-v1 (i.e., the models where accuracy for *unrelated* was not higher than the other labels), it is more likely that the gap originated from the strictness of group accuracy and the models' superficial understanding of will statements.

Further, roberta-large-mnli (trained with truncated inputs) showed better performance than longformer-base-4096 (trained with full length inputs). This suggests that models with token number limitations can still perform well with long inputs when truncated properly.

4.3 Experiment with Different Input Lengths

To investigate whether the models' performance was affected by input length and/or input truncation, we varied input lengths. Inputs were classified into three categories by length: a) short (equal or less than 192 or 256 tokens); b) regular (larger than 192 or 256 tokens and equal or less than 384 or 512); and c) long (larger than 384 or 512)[9]. The

[6] https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1

[7] https://huggingface.co/sentence-transformers/all-distilroberta-v1

[8] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

[9] For all-mpnet-base-v2 model, 192 and 384 were used as dividing points instead of 256 and 512, as its maximum token length is 384.

Exhibit 1026 Page 135

Model	Precision	Recall	F1	Accuracy	GA
Transformers models					
bert-base-uncased	81.47	82.20	81.81	87.06	49.02
distilbert-base-uncased	82.37	82.59	82.48	87.06	50.98
longformer-base-4096	94.09	93.69	93.85	94.90	74.51
roberta-large-mnli	**96.67**	**96.25**	**96.42**	**96.86**	**84.31**
Sentence-Transformers models					
all-MiniLM-L12-v2	81.50	84.54	82.64	85.49	50.98
all-disilroberta-v1	83.43	85.99	84.22	87.06	49.02
multi-qa-mpnet-base-dot-v1	91.48	94.13	92.65	93.73	74.51
all-mpnet-base-v2	92.93	95.05	93.82	94.90	76.47

Table 2: Results of all classifiers trained on our dataset using five measures: precision, recall, F1, accuracy, and group accuracy (GA in table, see Section 4.1). Precision, recall, and F1 scores were computed with Macro average. Group accuracy is calculated with positive and negative examples generated from the same statement as a unit. If a group contains one or more incorrect predictions, the group is considered incorrect.

Model	Precision	Recall	F1	Accuracy
all-mpnet-base-v2 (sample n = 52)				
token <= 192	93.75	93.15	92.91	94.23
192 < token <= 384	90.48	89.10	89.63	94.23
384 < token	86.35	89.49	87.08	88.46
longformer-base-4096 (sample n = 20)				
token <= 256	95.24	88.89	90.77	95.00
256 < token <= 512	80.56	71.67	70.93	80.00
512 < token	73.81	66.67	66.93	75.00
roberta-large-mnli (sample n = 20)				
token <= 256	100	100	100	100
256 < token <= 512	94.87	85.00	88.76	90.00
512 < token	94.44	93.33	93.27	95.00

Table 3: Results with different input lengths. Models generally performed worse when token numbers in inputs were larger. Roberta-large-mnli performed better than longformer-base-4096 in predicting long inputs (token > 512) despite truncation. Roberta-large-mnli also showed better performance on long inputs than regular inputs, indicating input truncation did not worsen the model's performance.

number of inputs in each category varied. To control the impact of varying sample size, categories with larger sample sizes were reduced by random sampling to match the smallest sample size. For this experiment, we used the three models with best performance on the full dataset.

Table 3 shows the results. Models generally performed worse with larger input token numbers. All three models performed best with inputs with smaller token numbers. One interesting finding is that roberta-large-mnli performed better than longformer-base-4096 in predicting inputs with large token numbers (n > 512) despite truncation. Roberta-large-mnli showed better performance on long inputs than regular inputs, indicating truncation did not negatively affect the model's performance. This finding aligns with the observation from the overall results that models with token number limitation can still show good performances on long inputs when properly truncated. However, results differed with a smaller token number limitation. All-mpnet-base-v2 performed worst on long inputs (where truncation occurred), indicating truncation negatively impacted performance.

4.4 Results from Ablation Experiment

To determine whether the models correctly rely on all three types of texts (statements, conditions, and laws), we conducted an ablation experiment, training the models with datasets lacking one (law or condition) or two types (law+condition) of inputs. Poliak et al. (2018) suggested a similar experiment

Exhibit 1026 Page 136

Model	Precision	Recall	F1	Accuracy	GA
Statements and laws					
all-mpnet-base-v2	75.64	76.47	75.98	81.96	27.45
longformer-base-4096	75.72	75.98	75.79	82.75	37.25
roberta-large-mnli	77.93	77.51	77.55	84.31	39.22
Statements and conditions					
all-mpnet-base-v2	46.67	57.13	44.80	44.31	0.00
longformer-base-4096	58.00	49.44	51.30	58.82	0.00
roberta-large-mnli	37.63	41.72	38.08	56.86	0.00
Statements only					
all-mpnet-base-v2	36.76	35.94	34.04	36.47	0.00
longformer-base-4096	18.17	33.33	23.52	54.51	0.00
roberta-large-mnli	18.17	33.33	23.52	54.51	0.00

Table 4: Results from the ablation experiment where three models were trained with datasets lacking one (law or condition) or two types (law+condition) of inputs. Results demonstrate that the models' performances deteriorate if one or more input type(s) is omitted. Thus, both types of inputs affect the models' performances, though laws have larger impact than conditions.

to this one to set hypothesis-only baselines. We used the three models with the best performances on the full dataset for this experiment.

Table 4 shows the results. Performance deteriorated if any input type was omitted. Among the models trained on partial data, those trained with statements and laws performed better, with F1 scores over 75%. However, group accuracy dropped to the 30s or even 20%s, indicating actual understanding decreased considerably. Models trained with statements and conditions performed significantly worse. Results ranged between 38% and 52% for F1, and group accuracy dropped to 0. Models trained only with statements performed worst, with F1 scores dropping to 35% or lower.

This degradation from partial data shows including both conditions and laws positively affects the models' performance. Models use all three types of inputs when making predictions[10]. Laws have more impact on performance than conditions, since deterioration in results without laws was more significant than from models trained without conditions. This is realistic, since laws set the parameters within which conditions operate to make a given provision valid or invalid.

4.5 Understanding Results with LIME

To clarify model behavior, we used Local Interpretable Model-Agnostic Explanations (LIME) to explain our classifiers' predictions (Ribeiro et al., 2016). We implemented LIME on predictions made by the best performing model (roberta-large-mnli, fine-tuned on our data). The LIME results revealed that the model tends to correctly rely on all three texts (statements, laws, conditions) for a prediction, and uses meaningful features in many cases. However, sometimes the model utilized features that are not intuitive for humans.

4.5.1 A Correct Example

Figure 3 shows a case where the model made a correct prediction based on features sensible to humans. Figure 3's text is a statement saying two or more witnesses witnessed the testator signing the will and signed in each other's presence. It includes a condition stating that one of the two witnesses was ineligible to serve under Tennessee law, and the law specifying witness eligibility in Tennessee. The given condition and the law invalidate the will statement, rendering it as *refute*. The model correctly predicted the outcome based on tokens such as *ineligible*, *witnesses*, *One*, *testament*, and *2*. The top two tokens with greatest impact on the model's prediction were *ineligible* and *witnesses*, with importance score of 0.60 and 0.14, respectively. It is

[10] A reviewer suggested an experiment where laws and conditions are packed into premises and treat the task as normal NLI. A pilot experiment with the best performing model (roberta-large-mnli) found a slight deterioration in performance when laws and conditions were combined as premises (precision: 93.09, recall: 93.89, F1: 93.41, accuracy: 94.12, GA: 74.51; see Table 2 for comparison with main result). The biggest difference came from GA score (74.51 vs. 84.31). This lowered GA score implies that the model's understanding worsened when conditions and laws were combined into premises. Given this preliminary result, it is likely that three-way distinction in input types positively contributes to the model's performance. Nevertheless, as the performance difference is not significantly large, further investigation is needed.

Exhibit 1026 Page 137

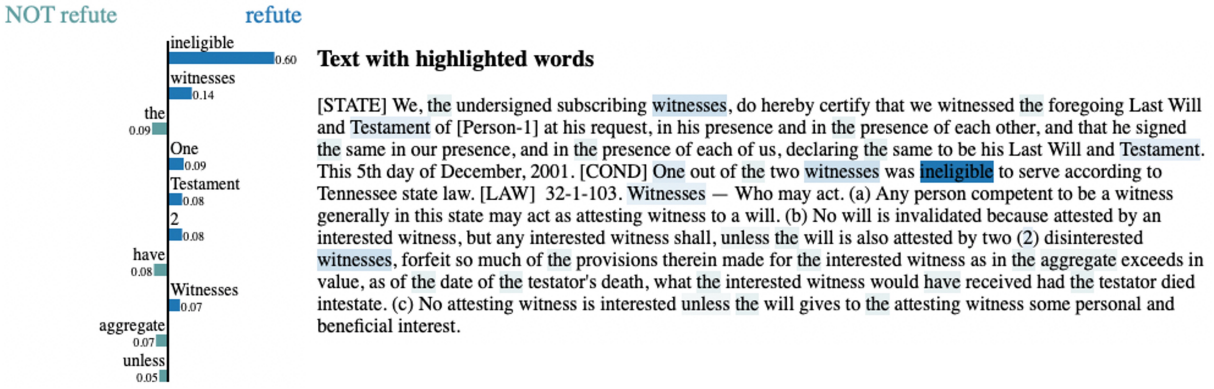


Figure 3: Examples of LIME explanations showing a case where the model (roberta-large-mnli) makes correct predictions based on features sensible to humans.

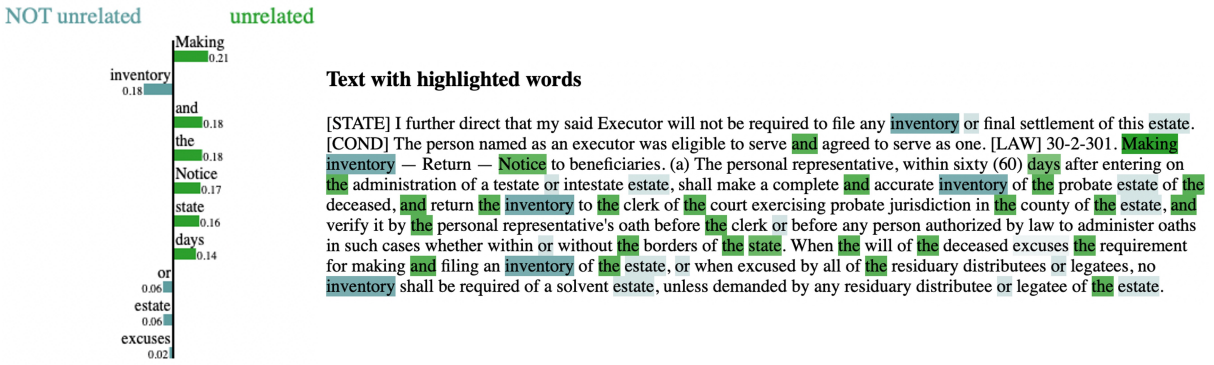


Figure 4: An example of a LIME explanation showing a case where the model (roberta-large-mnli) makes an incorrect prediction based on the tokens such as *Making*, *and*, and *the* which bear little semantic relevancy to the gist of the text.

understandable that these tokens have high scores, as they are the keywords ("**One** out of two (**2**) **witnesses** is **ineligible**") which provide grounds for revoking a will statement.

4.5.2 An Incorrect Example

Figure 4 shows a case where the model made an incorrect prediction based on irrelevant tokens. The intent of the will statement is to excuse the Executor from filing inventory. This is not contradicted by the condition and law, requiring the model to classify the text as *support*. Nevertheless, the model incorrectly classified it as *unrelated*. The reason for the incorrect prediction is that the model (probably) relied on irrelevant tokens such as *Making*, *and*, and *the*, which bear little semantic relevancy to the text, versus more relevant tokens such as *inventories* and *excuses* when making the prediction. This is likely due to overfitting induced by longer texts. This LIME explanation shows that even the best performing model still has room for improvement.

5 Future Work

This work can be expanded in several directions. First, it can be extended to cover multiple states by adapting the models or adding more data. Future work can investigate if models trained on a single state's data can be adapted to evaluate data from other states. Also, wills from multiple states could be added to the dataset. We expect the augmented dataset would enhance the models' performance on evaluating wills from other states.

This work can be expanded to other legal domains. Models trained on our data can be adapted to similar tasks in other legal domains.

Lastly, this work can be extended by investigating novel transformer models. Given the uniqueness of our dataset with regard to the number of inputs and text length, it is likely that further experiment and modification is needed to handle such characteristics.

Exhibit 1026 Page 138

6 Conclusion

This work presented an annotated dataset for natural language inference (NLI) in the legal domain, consisting of 1,014 data points generated from 23 publicly available wills. The dataset is novel for two reasons: it included texts with three input types (statement, law, and condition) rather than two (premise and hypothesis) in the traditional NLI, and included texts are longer than in general NLI datasets. The NLI models trained on our dataset showed reasonable performance in assessing the validity of will statements. Ablative experiments demonstrated that the models' performances worsen if any input type (condition, law, or both) is omitted. This suggests that the models utilize all three input types. The LIME implementation reveals that even the best performing model makes errors in some cases by using semantically irrelevant tokens. Our open-access dataset is publicly available at: https://github.com/ml4ai/nli4wills-corpus

Limitations

Our dataset consists of a relatively small number of data points (1,014 texts). Annotating will statements with relevant laws and conditions is a highly demanding, time intensive task. A larger size of dataset would likely further improve the models' performances.

Our dataset only includes wills executed and probated in Tennessee, with execution after 1970 and probate after 2000. Due to these restrictions, our framework might produce incorrect results on inputs from wills from different settings. Supplementing the dataset with wills from more diverse settings would address this limitation. Even though the scope is limited to a single state (Tennessee), our study demonstrates that transformer models trained on the dataset can evaluate the validity of statements from wills with reasonable accuracy. Future additions to our dataset will be available at the same URL: https://github.com/ml4ai/nli4wills-corpus

This work does not involve humans in the loop. Considering how crucial accuracy is for the task (i.e., legal validity evaluation), the work would have benefited much from involving domain experts. Even though our study discovered that the state-of-the-art transformer models can show good performances (over 85% accuracy in all 8 models) without human interaction, it was also found that the models' understanding on the task is rather superficial (GAs ranging from 49-85% in Table 2). Including humans-in-the-loop could be a solution for enhancing the models' understanding on the task.

Ethics Statement

We collected legal wills from Ancestry as a part of our dataset creation process. The wills probated in court are in the public domain in the US, and we did not violate Ancestry's Terms and Conditions when collecting wills. We also anonymized the wills by replacing any personal identifiable information contained in the documents with special tokens.

Our datasets and codes are released to public. We believe our released datasets and codes would contribute to society by promoting further NLI endeavors in the legal domain. The resources could potentially assist with people reviewing wills, but they should not be considered as legal advice. To avoid any confusion, we placed a disclaimer that the users must not rely on any information provided from our resources when making legal decisions and should instead consult with an attorney.

Acknowledgements

We thank the reviewers for their thoughtful comments and suggestions. This work was partially supported by the National Science Foundation (NSF) under grant 2217215, and by University of Arizona's Provost Investment Fund. Mihai Surdeanu and Clayton Morrison declare a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification,*

and Recognising Tectual Entailment, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

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.

Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. 2021. Back to square one: Artifact detection, training and commonsense disentanglement in the Winograd schema. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10486–10500, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering. *CoRR*, abs/2005.05257.

Yuta Koreeda and Christopher D. Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. *CoRR*, abs/2110.01799.

Xiao Liu, Da Yin, Yansong Feng, Yuting Wu, and Dongyan Zhao. 2021. Everything has a cause: Leveraging causal inference in legal text analysis. *CoRR*, abs/2104.09420.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Mobashir Sadat and Cornelia Caragea. 2022. SciNLI: A corpus for natural language inference on scientific text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7399–7409, Dublin, Ireland. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019a. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019b. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Sandeep Suntwal, Mithun Paul, Rebecca Sharp, and Mihai Surdeanu. 2019. On the importance of delexicalization for fact verification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3413–3418, Hong Kong, China. Association for Computational Linguistics.

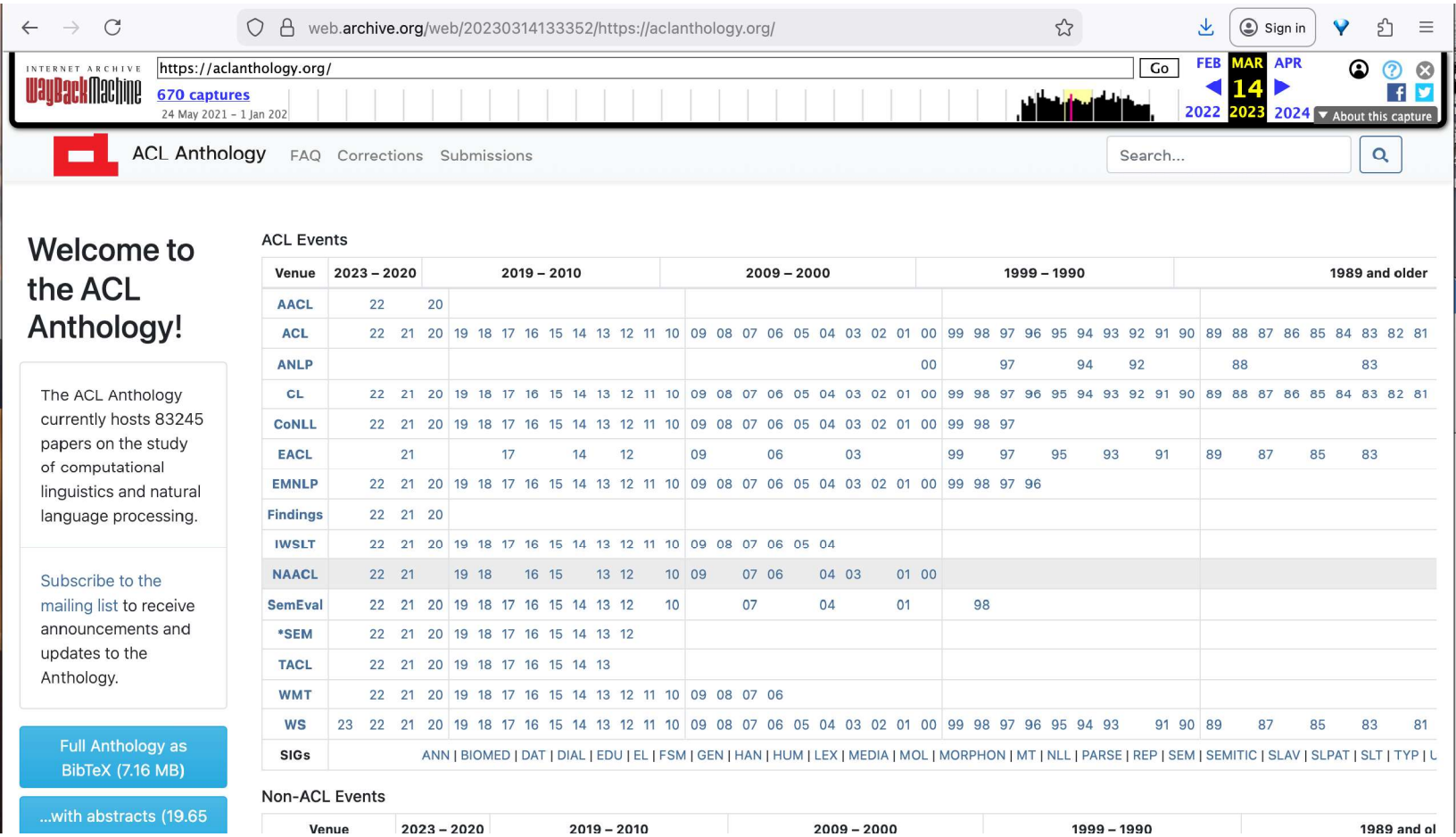
Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

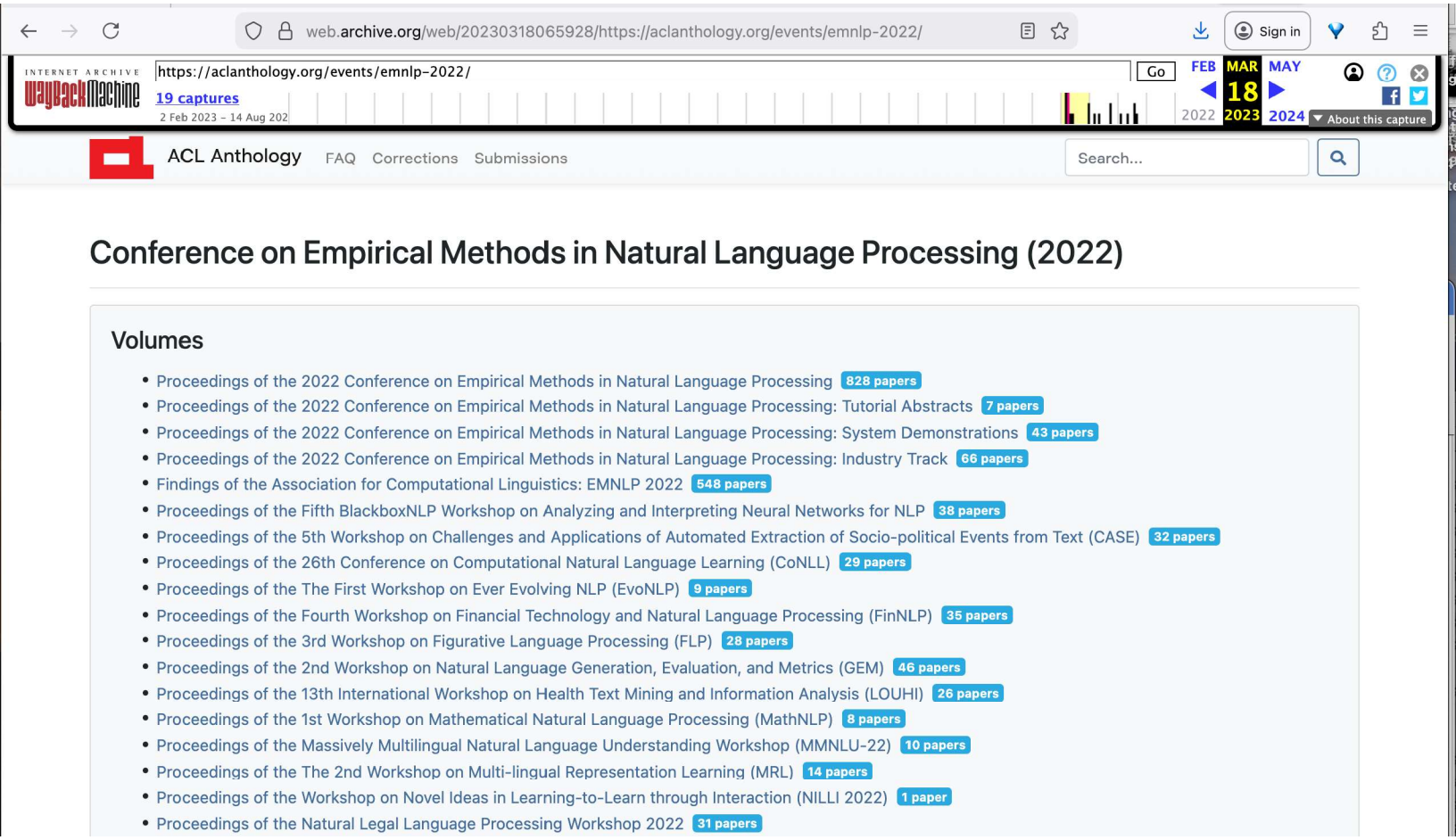
Ziyue Wang, Baoxin Wang, Xingyi Duan, Dayong Wu, Shijin Wang, Guoping Hu, and Ting Liu. 2019. IFlyLegal: A Chinese legal system for consultation, law searching, and document analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 97–102, Hong Kong, China. Association for Computational Linguistics.

Moshe Wasserblat. 2021. Sentence transformer fine-tuning (setfit): Outperforming gpt-3 on few-shot text-classification while being 1600 times smaller.

Exhibit 1026 Page 140

APPENDIX 1011E


web.archive.org/web/20230314133352/https://aclanthology.org/
Sign in
INTERNET ARCHIVE
WayBack Machine
https://aclanthology.org/
Go
670 captures
24 May 2021 - 1 Jan 202
FEB
MAR
14
APR
2022
2023
2024
About this capture
ACL Anthology
FAQ
Corrections
Submissions
Search...
Welcome to the ACL Anthology!
The ACL Anthology currently hosts 83245 papers on the study of computational linguistics and natural language processing.
Subscribe to the mailing list to receive announcements and updates to the Anthology.
Full Anthology as BibTeX (7.16 MB)
...with abstracts (19.65
ACL Events
Venue
2023 – 2020
2019 – 2010
2009 – 2000
1999 – 1990
1989 and older
AACL 22 20
ACL 22 21 20 19 18 17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00 99 98 97 96 95 94 93 92 91 90 89 88 87 86 85 84 83 82 81
ANLP 00 97 94 92 88 83
CL 22 21 20 19 18 17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00 99 98 97 96 95 94 93 92 91 90 89 88 87 86 85 84 83 82 81
CoNLL 22 21 20 19 18 17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00 99 98 97
EACL 21 17 14 12 09 06 03 99 97 95 93 91 89 87 85 83
EMNLP 22 21 20 19 18 17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00 99 98 97 96
Findings 22 21 20
IWSLT 22 21 20 19 18 17 16 15 14 13 12 11 10 09 08 07 06 05 04
NAACL 22 21 19 18 16 15 13 12 10 09 07 06 04 03 01 00
SemEval 22 21 20 19 18 17 16 15 14 13 12 10 07 04 01 98
*SEM 22 21 20 19 18 17 16 15 14 13 12
TACL 22 21 20 19 18 17 16 15 14 13
WMT 22 21 20 19 18 17 16 15 14 13 12 11 10 09 08 07 06
WS 23 22 21 20 19 18 17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00 99 98 97 96 95 94 93 91 90 89 87 85 83 81
SIGs
ANN | BIOMED | DAT | DIAL | EDU | EL | FSM | GEN | HAN | HUM | LEX | MEDIA | MOL | MORPHON | MT | NLL | PARSE | REP | SEM | SEMITIC | SLAV | SLPAT | SLT | TYP | U
Non-ACL Events
Venue
2023 – 2020
2019 – 2010
2009 – 2000
1999 – 1990
1989 and ol



Conference on Empirical Methods in Natural Language Processing (2022)

Volumes

- Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing 828 papers
- Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts 7 papers
- Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations 43 papers
- Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track 66 papers
- Findings of the Association for Computational Linguistics: EMNLP 2022 548 papers
- Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP 38 papers
- Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) 32 papers
- Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL) 29 papers
- Proceedings of the The First Workshop on Ever Evolving NLP (EvoNLP) 9 papers
- Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP) 35 papers
- Proceedings of the 3rd Workshop on Figurative Language Processing (FLP) 28 papers
- Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM) 46 papers
- Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI) 26 papers
- Proceedings of the 1st Workshop on Mathematical Natural Language Processing (MathNLP) 8 papers
- Proceedings of the Massively Multilingual Natural Language Understanding Workshop (MMNLU-22) 10 papers
- Proceedings of the The 2nd Workshop on Multi-lingual Representation Learning (MRL) 14 papers
- Proceedings of the Workshop on Novel Ideas in Learning-to-Learn through Interaction (NILLI 2022) 1 paper
- Proceedings of the Natural Legal Language Processing Workshop 2022 31 papers

web.archive.org/web/20230318065928/https://aclanthology.org/events/emnlp-2022/#2022findi
Sign in
INTERNET ARCHIVE WayBackMachine
https://aclanthology.org/events/emnlp-2022/
Go
19 captures
2 Feb 2023 - 14 Aug 202
FEB MAR MAY
2022 18 2023 2024
About this capture

pdf bib abs **Partitioned Gradient Matching-based Data Subset Selection for Compute-Efficient Robust ASR Training**
Ashish Mittal | Durga Sivasubramanian | Rishabh Iyer | Preethi Jyothi | Ganesh Ramakrishnan

pdf bib abs **Adaptive Graph Convolutional Network for Knowledge Graph Entity Alignment**
Renbo Zhu | Xukun Luo | Meng Ma | Ping Wang

pdf bib abs **Towards Robust NLG Bias Evaluation with Syntactically-diverse Prompts**
Arshiya Aggarwal | Jiao Sun | Nanyun Peng

pdf bib abs **PcMSP: A Dataset for Scientific Action Graphs Extraction from Polycrystalline Materials Synthesis Procedure Text**
Xianjun Yang | Ya Zhuo | Julia Zuo | Xinlu Zhang | Stephen Wilson | Linda Petzold

pdf bib abs **Validity Assessment of Legal Will Statements as Natural Language Inference**
Alice Kwak | Jacob Israelsen | Clayton Morrison | Derek Bambauer | Mihai Surdeanu

pdf bib abs **AdaPrompt: Adaptive Model Training for Prompt-based NLP**
Yulong Chen | Yang Liu | Li Dong | Shuohang Wang | Chenguang Zhu | Michael Zeng | Yue Zhang

pdf bib abs **Code Generation From Flowcharts with Texts: A Benchmark Dataset and An Approach**
Zejie Liu | Xiaoyu Hu | Deyu Zhou | Lin Li | Xu Zhang | Yanzheng Xiang

pdf bib abs **Focus! Relevant and Sufficient Context Selection for News Image Captioning**
Mingyang Zhou | Grace Luo | Anna Rohrbach | Zhou Yu

pdf bib abs **Generative Aspect-Based Sentiment Analysis with Contrastive Learning and Expressive Structure**
Joseph Peper | Lu Wang

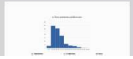
pdf bib abs **Semantic Dependency Parsing with Edge GNNs**
Songlin Yang | Kewei Tu

web.archive.org/web/20230202142404/https://aclanthology.org/2022.findings-emnlp.447.pdf
Sign in
INTERNET ARCHIVE Wayback Machine
https://aclanthology.org/2022.findings-emnlp.447.pdf
Go
2 captures
2 Feb 2023 - 26 Mar 2025
JAN FEB MAR
2022 02 2023 2024
About this capture
6076 (1 of 10)
Automatic Zoom









Validity Assessment of Legal Will Statements as Natural Language Inference

Alice Saebom Kwak, Jacob O. Israelsen, Clayton T. Morrison, Derek E. Bambauer, Mihai Surdeanu
The University of Arizona, Tucson, Arizona, USA
{alicekwak, jisraelsen, claytonm, derekbambauer, msurdeanu}@arizona.edu

Abstract

This work introduces a natural language inference (NLI) dataset that focuses on the validity of statements in legal wills. This dataset is unique because: (a) each entailment decision requires three inputs: the statement from the will, the law, and the conditions that hold at the time of the testator's death; and (b) the included texts are longer than the ones in current NLI datasets. We trained eight neural NLI models

NLI in general deals with determining whether a premise entails, contradicts, or is neutral to the hypothesis given. Our project made two important changes to this approach to fit legal documents. First, our dataset contains three input types: statements from wills (as hypotheses), laws (as premises), and *conditions*, which are circumstances at the time of probate (e.g., eligibility of the testator or witnesses). This adaptation is necessary

Validity Assessment of Legal Will Statements as Natural Language Inference

Alice Saebom Kwak, Jacob O. Israelsen, Clayton T. Morrison, Derek E. Bambauer, Mihai Surdeanu
The University of Arizona, Tucson, Arizona, USA
{alicekwak, jisraelsen, claytonm, derekbambauer, msurdeanu}@arizona.edu

Abstract

This work introduces a natural language inference (NLI) dataset that focuses on the validity of statements in legal wills. This dataset is unique because: (a) each entailment decision requires three inputs: the statement from the will, the law, and the conditions that hold at the time of the testator's death; and (b) the included texts are longer than the ones in current NLI datasets. We trained eight neural NLI models in this dataset. All the models achieve more than 80% macro F1 and accuracy, which indicates that neural approaches can handle this task reasonably well. However, group accuracy, a stricter evaluation measure that is calculated with a group of positive and negative examples generated from the same statement as a unit, is in mid 80s at best, which suggests that the models' understanding of the task remains superficial. Further ablative analyses and explanation experiments indicate that all three text segments are used for prediction, but some decisions rely on semantically irrelevant tokens. This indicates that overfitting on these longer texts likely happens, and that additional research is required for this task to be solved.

1 Introduction

Natural language inference (NLI) in the legal domain has not been widely investigated, despite its importance and potential. One such important NLI application is validity assessment of legal documents such as wills. Legal procedures for creating and executing wills are evolving rapidly. Processing a will via probate is a costly, time-consuming process that can be exacerbated by errors or by challenges to validity. These problems will likely increase as people increasingly employ electronic wills. Developing natural language techniques that can determine a will's validity at creation and execution can increase validity, conserve legal resources, and effectuate the author's intent. To this end, this work explores how NLI models can be employed to evaluate the validity of will statements.

NLI in general deals with determining whether a premise entails, contradicts, or is neutral to the hypothesis given. Our project made two important changes to this approach to fit legal documents. First, our dataset contains three input types: statements from wills (as hypotheses), laws (as premises), and *conditions*, which are circumstances at the time of probate (e.g., eligibility of the testator or witnesses). This adaptation is necessary as will statements' validity cannot be evaluated without considering both circumstances at the time of probate ("conditions") and relevant legal rules ("laws"). Also, we switched the labels from entailment, contradict, neutral to support, refute, unrelated to better represent the relation between will statements and laws.
The major contributions of this work are:

- We create an open-access annotated dataset with 1,014 data points, generated from 23 publicly available wills. In addition to the three-tuple setting unique to the legal domain, this dataset also contains texts considerably longer than in other open-domain NLI datasets, such as SNLI (Bowman et al., 2015). The average length of our texts is 269 tokens, while the average lengths of premises and hypotheses in SNLI are 8 and 14 tokens, respectively.

- We demonstrate that validity assessment of legal will statements can be handled reasonably well with state-of-the-art NLI models when trained on our dataset. However, low scores in group accuracy, which is a stricter evaluation measure calculated with a group of positive and negative examples generated from the same statement as a unit, indicate that work remains before NLI models fully understand legal language.

- We explain how the trained models utilize our dataset via ablation tests that remove various

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6076 - 6085
December 7-11, 2022 ©2022 Association for Computational Linguistics
Exhibit 1026 Page 146

pieces of information at prediction time (laws, conditions, or both), and through post-hoc explainability analyses using LIME (Ribeiro et al., 2016). Our analyses indicate that in most cases the NLI classifiers use meaningful phrases from all three pieces of text, indicating the models do capture useful information. However, in some situations, the models use features that are not intuitive for humans, indicating that the task is incomplete.

2 Related Work

Natural language inference (NLI), also known as Recognizing Textual Entailment (RTE), determines entailment relations between a pair of sentences: a premise and a hypothesis. Their relationship is either: a) entailment, if a premise entails a hypothesis; b) contradiction, if a premise contradicts a hypothesis; or c) neutral, if a premise neither entails nor contradicts a hypothesis. NLI has been a key framework in natural language processing since Dagan et al. (2006) proposed the RTE challenge.

Numerous datasets exist for general NLI tasks. For example, the Stanford Natural Language Inference (SNLI) Corpus (Bowman et al., 2015) dataset contains about 570,000 pairs of sentences (premises and hypotheses) generated from Flickr image captions. More recently, datasets for domain-specific NLI tasks are being introduced. SciNLI (Sadat and Caragea, 2022) is drawn from scientific texts. It contains 107,412 sentence pairs extracted from papers in natural language processing and computational linguistics.

There have been a few NLI-related resources for the legal domain, including IFlyLegal (Wang et al., 2019), StAtutory Reasoning Assessment (SARA) dataset (Holzenberger et al., 2020), Graph-based Causal Inference (GCI) framework (Liu et al., 2021), and ContractNLI (Koreeda and Manning, 2021). One IFlyLegal module was developed for natural language article inference; it provides relevant legal articles when asked legal questions. SARA contains rules extracted from the US Internal Revenue Code and natural language questions. GCI generates causal graphs based on fact descriptions. Though the framework was not specifically designed for law, Liu et al. (2021) demonstrated that it can be utilized for legal text analysis. ContractNLI is a dataset consisting of 607 legal contracts designed to handle document-level natural language inference.

Our work differs from these existing NLI works in two ways: (a) it contains three types of input information; and (b) it operates in a pragmatic middle space with respect to text length: larger than datasets with sentence-level texts, which are insufficient to capture legal details, but shorter than document-level inference datasets.

3 Dataset

Our dataset includes three types of entries (will statements, applicable law, and facts representing external state) rather than the usual two types. Validity often depends upon such external circumstances. Thus, a will statement can be legally valid under some facts, but not others. Hence, a will can change from valid to invalid or vice versa; wills must, at minimum, be evaluated when executed (drafted and signed) and when probated (where a court determines whether to implement its provisions). For example, if a will contains a statement appointing a specific person as executor, that person's eligibility must be verified. This requires information about both the law and the person's circumstances at probate time. For example, according to Tennessee Code section 40-20-115, any person who has been imprisoned cannot be an executor. Thus, to assess eligibility of a Tennessee executor in Tennessee, one must know whether they have been in prison at probate time.

3.1 Data Collection

Our data was collected from the U.S. Wills and Probates datasets in Ancestry, which contains documents in the public domain.[1] We chose 23 wills based on three criteria: 1) whether the wills were handwritten; 2) the wills' execution and probation dates; and 3) where the wills were executed and probated. We excluded handwritten wills due to the difficulty of OCR text recognition. Execution and probation dates can affect interpretation of wills; we excluded wills executed before 1970 and probated before 2000. Lastly, we only collected wills from Tennessee, because including wills from multiple states would require analyzing each state's laws governing wills. Tennessee had the greatest number of wills meeting our criteria.

All collected wills were anonymized. Personal information was replaced with special tokens denoting the type of information (such as [Person-

[1]Court documents are in the public domain in the US.

Exhibit 1026 Page 147

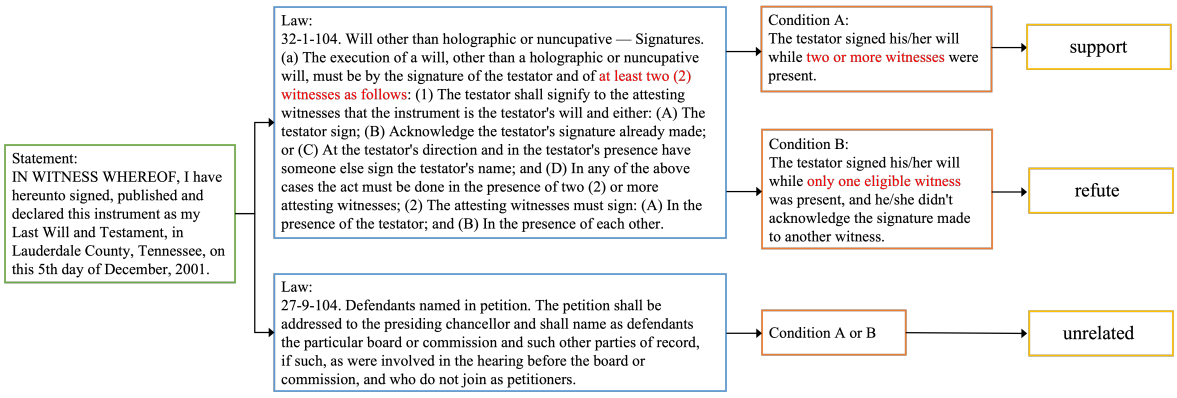


Figure 1: A demonstration of the annotation procedure. Given a law and condition, each statement is evaluated for validity. *Support*: it is supported by the law and condition provided; *Refute*: refuted by the law and condition. *Unrelated*: not relevant to the law. When a statement and a law are unrelated to each other, a condition is not required for classification. However, it is still assigned one of the conditions from either support or refute case. The goal is to prevent NLI models from depending on contextual difference (i.e., "texts without conditions are all unrelated cases") rather than features when making predictions.

n], [Address-n], [Number-n], where n identified each person or object in the will. (Suntwal et al. (2019) suggested this anonymization method). Anonymization was performed manually to prevent including personal information in the dataset.

3.2 Annotation

Each will statement was annotated as *support*, *refute*, or *unrelated* based on a given condition and law. (Our annotators set hypothetical conditions when necessary to categorize statements.)

Importantly, each statement was annotated multiple times. A statement was annotated once as support, once as refute, and thrice as unrelated.[2] This ensured that every statement was labeled with all three classes in the same ratio (support:refute:unrelated=1:1:3). Also, annotating statements three times more with "unrelated" than support and refute enabled including a greater range of laws in the dataset. Formally, our annotation procedure included the following steps:

(1) preprocessing: extracting texts from collected wills using OCR[3], and copying statements into the dataset;

(2) for each statement, identifying the Tennessee legal provision that supported, refuted, or was unrelated to the statement's validity;

(3) adding a condition specifying external circumstances relevant to validity;

(4) repeating these steps to generate five annotations per statement;

(5) anonymizing all statements by replacing personal information with special tokens.

Figure 1 demonstrates the annotation procedure with examples.

Two annotators participated in the task. One is a law student, and annotation was supervised by author Bambauer, a law professor. The other annotator does not have legal training, but ongoing discussions and reviews ensured the uniform quality. The annotators contributed equally. Annotators worked individually but shared annotation guidelines. After the annotation was complete, cross-annotation was conducted on 200 randomly selected items (100 items from each annotator's set; each annotator worked blinded on items drawn from the other annotator's set) to check inter-annotator agreement. We calculated Kappa agreement score based on the cross-annotation result, and the score is 0.91 (rounded to hundredth).

After the annotation was complete, the dataset was split into training, development, and test sets. There are 1014 data points in our dataset, split 50:25:25% (train: 504, development: 255, test: 255). When splitting, a group containing annotations from a single statement (one support, one refute, and three unrelated) was treated as a single

[2] A few exceptions occurred when:
a) it was impossible to create a case where the statement was refuted. In such cases, statements were not annotated as "refute."
b) a statement had more than one relevant legal rule. Such statements were annotated twice as support, twice as refute, and once as unrelated.

[3] The collected wills were scanned documents, so OCR was necessary to extract machine-readable text.

Exhibit 1026 Page 148

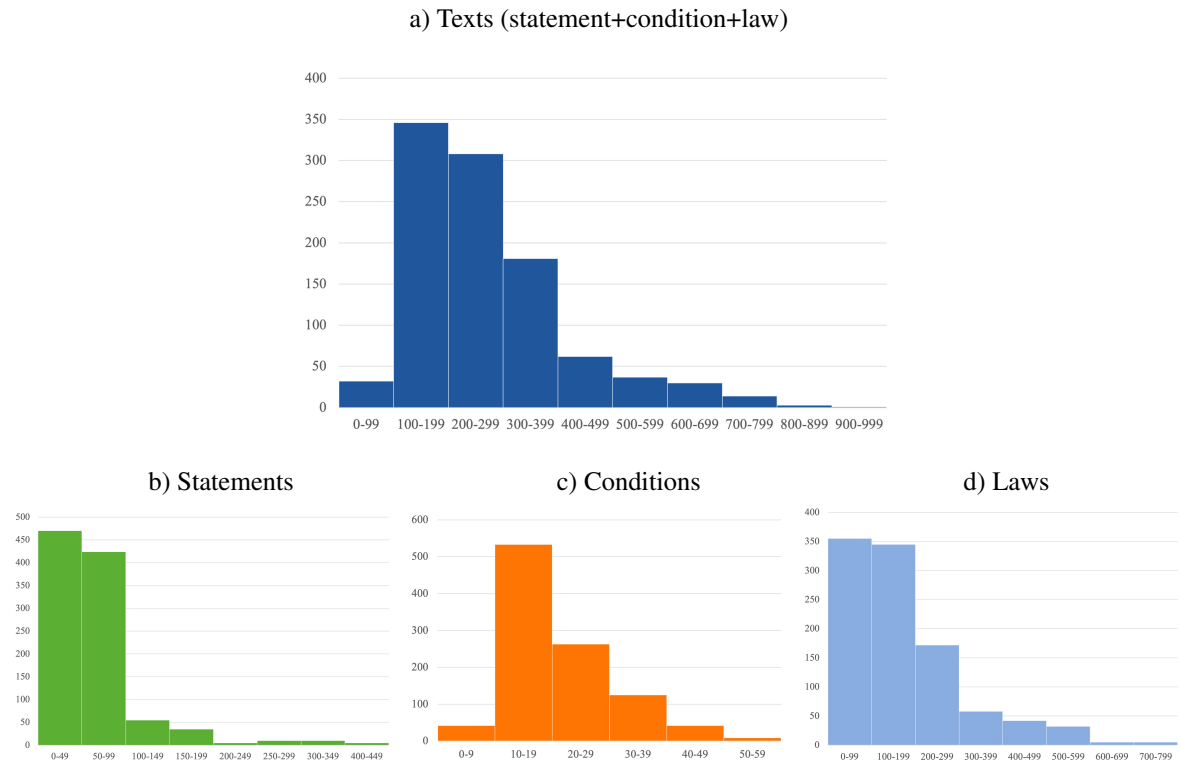


Figure 2: Histograms demonstrating the length distribution of a) full texts (i.e., statement+condition+law), b) statements, c) conditions, and d) laws in terms of token counts. Token counts are plotted on the X-axis; the Y-axis indicates the number of texts/statements/conditions/laws for each bin.

Datasets	**Texts**	**Tokens**
Train	504	135218
Development	255	69643
Test	255	68042
Total	**1014**	**272903**

Table 1: The number of texts and tokens contained in each dataset and in total

unit. Each annotator's work was equally represented in all three datasets. All 23 wills appear in more than one of the train, dev, and test partitions. However, given the nature of the will statements from the wills and how they were annotated, data leakage is unlikely. Will statements are independent and generally do not convey information about other statements from the same wills. Summary statistics are in Table 1 and Figure 2.

3.3 Characteristics of the Dataset

Our dataset has two characteristics distinguishing it from other datasets: a) texts are composed of three input types, and b) texts contain a large number of tokens.

First, our task requires a third input type: a condition, since validity often depends upon conditions.

Second, our texts tend to contain a large number of tokens. 44% of our texts contain 200 or more tokens. The average token number was 269.14; 79 texts contain more than 512 tokens (the threshold for most NLI models). Popular NLI datasets are shorter.

3.4 Implementation

We trained multiple NLI models with our dataset to assess their performance: four transformer models and four sentence-transformer models. The transformer models include bert-base-uncased (Devlin et al., 2018), distilbert-base-uncased (Sanh et al., 2019a), roberta-large-mnli (Liu et al., 2019), and longformer-base-4096 (Beltagy et al., 2020). Bert-base-uncased and dilstilbert-base-uncased were trained to set baselines on the task. In addition to the baseline models, we used roberta-large-mnli and longformer-base-4096. Among sentence-transformer models (Reimers and Gurevych, 2019), four pretrained models with top average performances (based on the Model Overview page[4]) were chosen: all-mpnet-base-v2[5], multi-qa-mpnet-base-

[4] https://www.sbert.net

[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Exhibit 1026 Page 149

dot-v1[6], all-distilroberta-v1 (Sanh et al., 2019b)[7], and all-MiniLM-L12-v2 (Wang et al., 2020)[8].

To distinguish between statements, laws, and conditions in the concatenated texts, we prefixed each with a special token: `[STATE]`, `[LAW]`, and `[COND]`.

Both transformer models and sentence-transformer models were trained on PyTorch 1.11.0 with Cuda 11.3. All transformer models were trained using the Trainer class provided by HuggingFace. Learning rates and training epochs were tuned on the development partition. Sentence-transformer models were trained with Sentence Transformer Fine-Tuning (SetFit) proposed by Wasserblat (2021). It was slightly adapted to fit the multi-class classification task, but its fundamentals remain intact. SetFit utilizes sentence pairs generated within the same class as training data. It fits the model with the data to minimize the Softmax Loss. Once the model is fitted, it is used to encode the training and development (or test, when it is a testing phase) datasets. The encoded data is used to fit the Logistic Regression classifier (with 'liblinear' solver) which, finally, makes predictions.

All models except for longformer-base-4096 were trained and evaluated on the truncated datasets due to the models' token number limitations. The models can only process texts with 512 or fewer tokens without truncation (all-MiniLM-L12-v2: 256 tokens; all-mpnet-base-v2: 384 tokens; other models, except longformer-base-4096: 512 tokens). When truncating datasets, the ratio of average token numbers among three input types (statements, laws, and conditions) was considered to prevent losing excessive information from a single type of input.

4 Results and Analysis

4.1 Evaluation Measures

We report results using standard accuracy, precision, recall, and F1 scores. To ensure all labels are well represented, precision, recall, and F1 scores were computed with Macro average. Additionally, as suggested by Elazar et al. (2021), we introduce a new measure called group accuracy (GA). GA is calculated with a group of positive and negative examples from the same statement (rather than each text) as a unit. If a group has one or more incorrect predictions, the group is incorrect. If a model correctly understands a will statement, it should perform equally well on all examples from the same statement.

4.2 Results from Trained Models

Table 2 shows the models' performances on our test partition. Overall, the models demonstrated good performances. Each achieved more than 80% in all metrics but group accuracy. Roberta-large-mnli showed the best performance. It achieved over 96% in all metrics but group accuracy (84.31%), suggesting it can handle the task reasonably well. However, the large difference between accuracy and group accuracy suggests that the models' understanding of the task is superficial.

Accuracy for the *unrelated* label was higher than for *support* and *refute* in all models except all-mpnet-base-v2 and multi-qa-mpnet-base-dot-v1. This higher accuracy for *unrelated* label may be partially attributable to the gap between accuracy and group accuracy, as it would inflate overall accuracy. However, since the gap between accuracy and group accuracy is not significantly smaller for all-mpnet-base-v2 and multi-qa-mpnet-base-dot-v1 (i.e., the models where accuracy for *unrelated* was not higher than the other labels), it is more likely that the gap originated from the strictness of group accuracy and the models' superficial understanding of will statements.

Further, roberta-large-mnli (trained with truncated inputs) showed better performance than longformer-base-4096 (trained with full length inputs). This suggests that models with token number limitations can still perform well with long inputs when truncated properly.

4.3 Experiment with Different Input Lengths

To investigate whether the models' performance was affected by input length and/or input truncation, we varied input lengths. Inputs were classified into three categories by length: a) short (equal or less than 192 or 256 tokens); b) regular (larger than 192 or 256 tokens and equal or less than 384 or 512); and c) long (larger than 384 or 512)[9]. The

[6] https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1

[7] https://huggingface.co/sentence-transformers/all-distilroberta-v1

[8] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

[9] For all-mpnet-base-v2 model, 192 and 384 were used as dividing points instead of 256 and 512, as its maximum token length is 384.

Exhibit 1026 Page 150

Model	Precision	Recall	F1	Accuracy	GA
Transformers models					
bert-base-uncased	81.47	82.20	81.81	87.06	49.02
distilbert-base-uncased	82.37	82.59	82.48	87.06	50.98
longformer-base-4096	94.09	93.69	93.85	94.90	74.51
roberta-large-mnli	**96.67**	**96.25**	**96.42**	**96.86**	**84.31**
Sentence-Transformers models					
all-MiniLM-L12-v2	81.50	84.54	82.64	85.49	50.98
all-disilroberta-v1	83.43	85.99	84.22	87.06	49.02
multi-qa-mpnet-base-dot-v1	91.48	94.13	92.65	93.73	74.51
all-mpnet-base-v2	92.93	95.05	93.82	94.90	76.47

Table 2: Results of all classifiers trained on our dataset using five measures: precision, recall, F1, accuracy, and group accuracy (GA in table, see Section 4.1). Precision, recall, and F1 scores were computed with Macro average. Group accuracy is calculated with positive and negative examples generated from the same statement as a unit. If a group contains one or more incorrect predictions, the group is considered incorrect.

Model	Precision	Recall	F1	Accuracy
all-mpnet-base-v2 (sample n = 52)				
token <= 192	93.75	93.15	92.91	94.23
192 < token <= 384	90.48	89.10	89.63	94.23
384 < token	86.35	89.49	87.08	88.46
longformer-base-4096 (sample n = 20)				
token <= 256	95.24	88.89	90.77	95.00
256 < token <= 512	80.56	71.67	70.93	80.00
512 < token	73.81	66.67	66.93	75.00
roberta-large-mnli (sample n = 20)				
token <= 256	100	100	100	100
256 < token <= 512	94.87	85.00	88.76	90.00
512 < token	94.44	93.33	93.27	95.00

Table 3: Results with different input lengths. Models generally performed worse when token numbers in inputs were larger. Roberta-large-mnli performed better than longformer-base-4096 in predicting long inputs (token > 512) despite truncation. Roberta-large-mnli also showed better performance on long inputs than regular inputs, indicating input truncation did not worsen the model's performance.

number of inputs in each category varied. To control the impact of varying sample size, categories with larger sample sizes were reduced by random sampling to match the smallest sample size. For this experiment, we used the three models with best performance on the full dataset.

Table 3 shows the results. Models generally performed worse with larger input token numbers. All three models performed best with inputs with smaller token numbers. One interesting finding is that roberta-large-mnli performed better than longformer-base-4096 in predicting inputs with large token numbers (n > 512) despite truncation. Roberta-large-mnli showed better performance on long inputs than regular inputs, indicating truncation did not negatively affect the model's performance. This finding aligns with the observation from the overall results that models with token number limitation can still show good performances on long inputs when properly truncated. However, results differed with a smaller token number limitation. All-mpnet-base-v2 performed worst on long inputs (where truncation occurred), indicating truncation negatively impacted performance.

4.4 Results from Ablation Experiment

To determine whether the models correctly rely on all three types of texts (statements, conditions, and laws), we conducted an ablation experiment, training the models with datasets lacking one (law or condition) or two types (law+condition) of inputs. Poliak et al. (2018) suggested a similar experiment

Exhibit 1026 Page 151

Model	Precision	Recall	F1	Accuracy	GA
Statements and laws					
all-mpnet-base-v2	75.64	76.47	75.98	81.96	27.45
longformer-base-4096	75.72	75.98	75.79	82.75	37.25
roberta-large-mnli	77.93	77.51	77.55	84.31	39.22
Statements and conditions					
all-mpnet-base-v2	46.67	57.13	44.80	44.31	0.00
longformer-base-4096	58.00	49.44	51.30	58.82	0.00
roberta-large-mnli	37.63	41.72	38.08	56.86	0.00
Statements only					
all-mpnet-base-v2	36.76	35.94	34.04	36.47	0.00
longformer-base-4096	18.17	33.33	23.52	54.51	0.00
roberta-large-mnli	18.17	33.33	23.52	54.51	0.00

Table 4: Results from the ablation experiment where three models were trained with datasets lacking one (law or condition) or two types (law+condition) of inputs. Results demonstrate that the models' performances deteriorate if one or more input type(s) is omitted. Thus, both types of inputs affect the models' performances, though laws have larger impact than conditions.

to this one to set hypothesis-only baselines. We used the three models with the best performances on the full dataset for this experiment.

Table 4 shows the results. Performance deteriorated if any input type was omitted. Among the models trained on partial data, those trained with statements and laws performed better, with F1 scores over 75%. However, group accuracy dropped to the 30s or even 20%s, indicating actual understanding decreased considerably. Models trained with statements and conditions performed significantly worse. Results ranged between 38% and 52% for F1, and group accuracy dropped to 0. Models trained only with statements performed worst, with F1 scores dropping to 35% or lower.

This degradation from partial data shows including both conditions and laws positively affects the models' performance. Models use all three types of inputs when making predictions[10]. Laws have more impact on performance than conditions, since deterioration in results without laws was more significant than from models trained without conditions. This is realistic, since laws set the parameters within which conditions operate to make a given provision valid or invalid.

4.5 Understanding Results with LIME

To clarify model behavior, we used Local Interpretable Model-Agnostic Explanations (LIME) to explain our classifiers' predictions (Ribeiro et al., 2016). We implemented LIME on predictions made by the best performing model (roberta-large-mnli, fine-tuned on our data). The LIME results revealed that the model tends to correctly rely on all three texts (statements, laws, conditions) for a prediction, and uses meaningful features in many cases. However, sometimes the model utilized features that are not intuitive for humans.

4.5.1 A Correct Example

Figure 3 shows a case where the model made a correct prediction based on features sensible to humans. Figure 3's text is a statement saying two or more witnesses witnessed the testator signing the will and signed in each other's presence. It includes a condition stating that one of the two witnesses was ineligible to serve under Tennessee law, and the law specifying witness eligibility in Tennessee. The given condition and the law invalidate the will statement, rendering it as *refute*. The model correctly predicted the outcome based on tokens such as *ineligible*, *witnesses*, *One*, *testament*, and *2*. The top two tokens with greatest impact on the model's prediction were *ineligible* and *witnesses*, with importance score of 0.60 and 0.14, respectively. It is

[10] A reviewer suggested an experiment where laws and conditions are packed into premises and treat the task as normal NLI. A pilot experiment with the best performing model (roberta-large-mnli) found a slight deterioration in performance when laws and conditions were combined as premises (precision: 93.09, recall: 93.89, F1: 93.41, accuracy: 94.12, GA: 74.51; see Table 2 for comparison with main result). The biggest difference came from GA score (74.51 vs. 84.31). This lowered GA score implies that the model's understanding worsened when conditions and laws were combined into premises. Given this preliminary result, it is likely that three-way distinction in input types positively contributes to the model's performance. Nevertheless, as the performance difference is not significantly large, further investigation is needed.

Exhibit 1026 Page 152

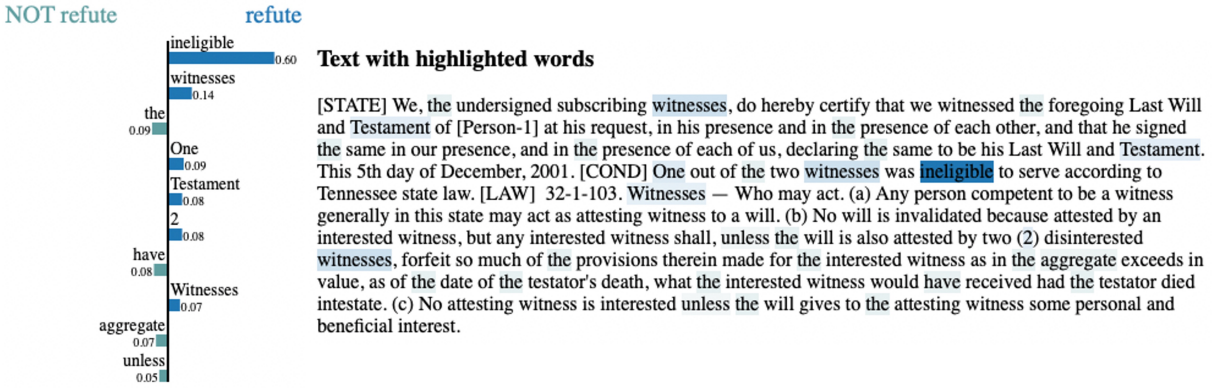


Figure 3: Examples of LIME explanations showing a case where the model (roberta-large-mnli) makes correct predictions based on features sensible to humans.

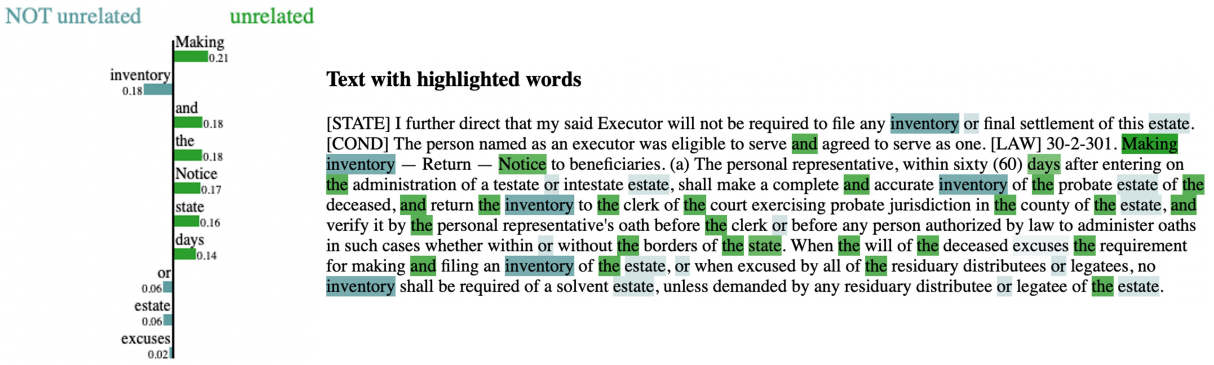


Figure 4: An example of a LIME explanation showing a case where the model (roberta-large-mnli) makes an incorrect prediction based on the tokens such as *Making*, *and*, and *the* which bear little semantic relevancy to the gist of the text.

understandable that these tokens have high scores, as they are the keywords ("**One** out of two (**2**) **witnesses** is **ineligible**") which provide grounds for revoking a will statement.

4.5.2 An Incorrect Example

Figure 4 shows a case where the model made an incorrect prediction based on irrelevant tokens. The intent of the will statement is to excuse the Executor from filing inventory. This is not contradicted by the condition and law, requiring the model to classify the text as *support*. Nevertheless, the model incorrectly classified it as *unrelated*. The reason for the incorrect prediction is that the model (probably) relied on irrelevant tokens such as *Making*, *and*, and *the*, which bear little semantic relevancy to the text, versus more relevant tokens such as *inventories* and *excuses* when making the prediction. This is likely due to overfitting induced by longer texts. This LIME explanation shows that even the best performing model still has room for improvement.

5 Future Work

This work can be expanded in several directions. First, it can be extended to cover multiple states by adapting the models or adding more data. Future work can investigate if models trained on a single state's data can be adapted to evaluate data from other states. Also, wills from multiple states could be added to the dataset. We expect the augmented dataset would enhance the models' performance on evaluating wills from other states.

This work can be expanded to other legal domains. Models trained on our data can be adapted to similar tasks in other legal domains.

Lastly, this work can be extended by investigating novel transformer models. Given the uniqueness of our dataset with regard to the number of inputs and text length, it is likely that further experiment and modification is needed to handle such characteristics.

Exhibit 1026 Page 153

6 Conclusion

This work presented an annotated dataset for natural language inference (NLI) in the legal domain, consisting of 1,014 data points generated from 23 publicly available wills. The dataset is novel for two reasons: it included texts with three input types (statement, law, and condition) rather than two (premise and hypothesis) in the traditional NLI, and included texts are longer than in general NLI datasets. The NLI models trained on our dataset showed reasonable performance in assessing the validity of will statements. Ablative experiments demonstrated that the models' performances worsen if any input type (condition, law, or both) is omitted. This suggests that the models utilize all three input types. The LIME implementation reveals that even the best performing model makes errors in some cases by using semantically irrelevant tokens. Our open-access dataset is publicly available at: https://github.com/ml4ai/nli4wills-corpus

Limitations

Our dataset consists of a relatively small number of data points (1,014 texts). Annotating will statements with relevant laws and conditions is a highly demanding, time intensive task. A larger size of dataset would likely further improve the models' performances.

Our dataset only includes wills executed and probated in Tennessee, with execution after 1970 and probate after 2000. Due to these restrictions, our framework might produce incorrect results on inputs from wills from different settings. Supplementing the dataset with wills from more diverse settings would address this limitation. Even though the scope is limited to a single state (Tennessee), our study demonstrates that transformer models trained on the dataset can evaluate the validity of statements from wills with reasonable accuracy. Future additions to our dataset will be available at the same URL: https://github.com/ml4ai/nli4wills-corpus

This work does not involve humans in the loop. Considering how crucial accuracy is for the task (i.e., legal validity evaluation), the work would have benefited much from involving domain experts. Even though our study discovered that the state-of-the-art transformer models can show good performances (over 85% accuracy in all 8 models) without human interaction, it was also found that the models' understanding on the task is rather superficial (GAs ranging from 49-85% in Table 2). Including humans-in-the-loop could be a solution for enhancing the models' understanding on the task.

Ethics Statement

We collected legal wills from Ancestry as a part of our dataset creation process. The wills probated in court are in the public domain in the US, and we did not violate Ancestry's Terms and Conditions when collecting wills. We also anonymized the wills by replacing any personal identifiable information contained in the documents with special tokens.

Our datasets and codes are released to public. We believe our released datasets and codes would contribute to society by promoting further NLI endeavors in the legal domain. The resources could potentially assist with people reviewing wills, but they should not be considered as legal advice. To avoid any confusion, we placed a disclaimer that the users must not rely on any information provided from our resources when making legal decisions and should instead consult with an attorney.

Acknowledgements

We thank the reviewers for their thoughtful comments and suggestions. This work was partially supported by the National Science Foundation (NSF) under grant 2217215, and by University of Arizona's Provost Investment Fund. Mihai Surdeanu and Clayton Morrison declare a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification,*

and Recognising Tectual Entailment, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

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.

Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. 2021. Back to square one: Artifact detection, training and commonsense disentanglement in the Winograd schema. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10486–10500, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering. *CoRR*, abs/2005.05257.

Yuta Koreeda and Christopher D. Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. *CoRR*, abs/2110.01799.

Xiao Liu, Da Yin, Yansong Feng, Yuting Wu, and Dongyan Zhao. 2021. Everything has a cause: Leveraging causal inference in legal text analysis. *CoRR*, abs/2104.09420.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Mobashir Sadat and Cornelia Caragea. 2022. SciNLI: A corpus for natural language inference on scientific text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7399–7409, Dublin, Ireland. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019a. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019b. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

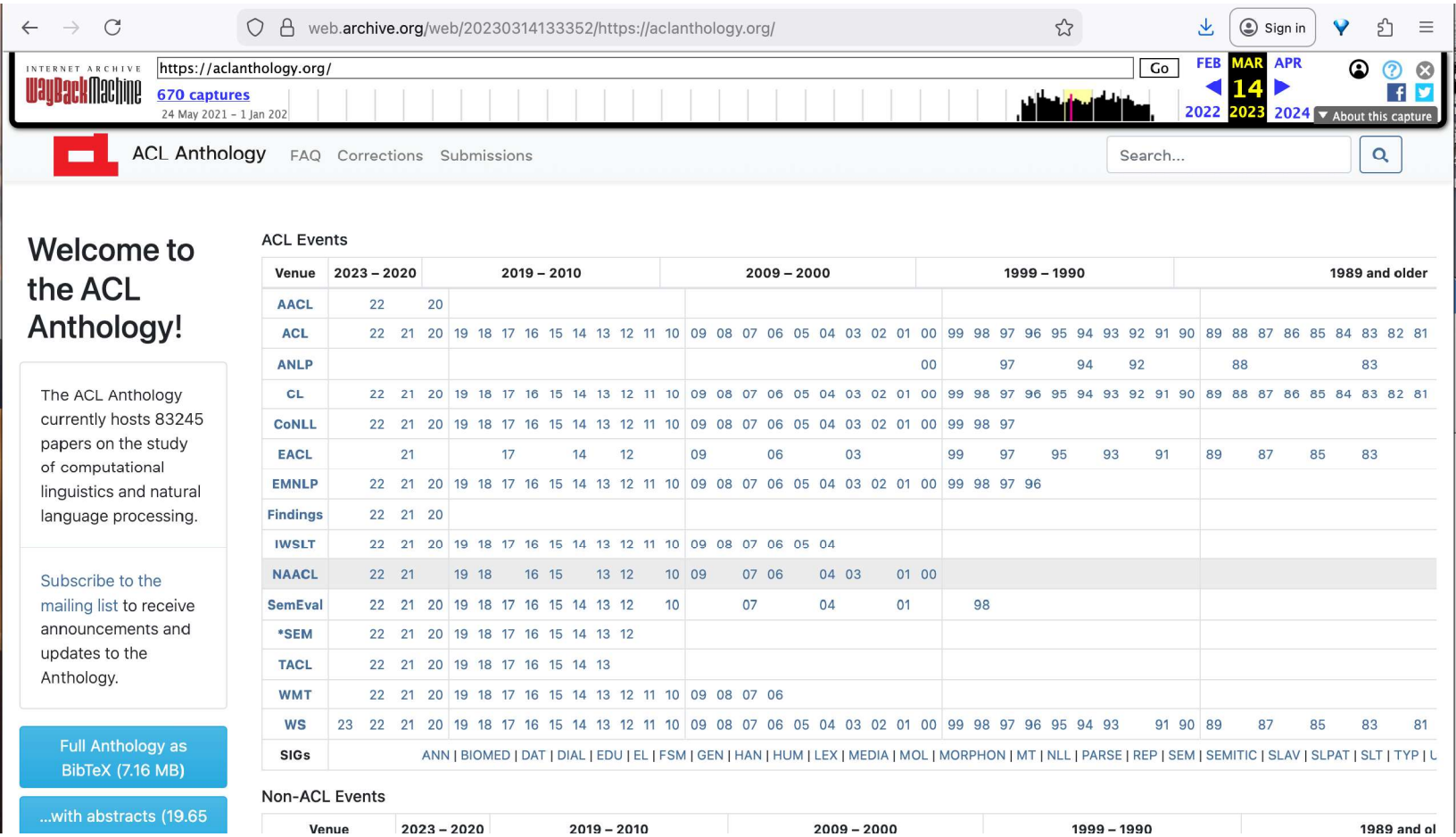
Sandeep Suntwal, Mithun Paul, Rebecca Sharp, and Mihai Surdeanu. 2019. On the importance of delexicalization for fact verification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3413–3418, Hong Kong, China. Association for Computational Linguistics.

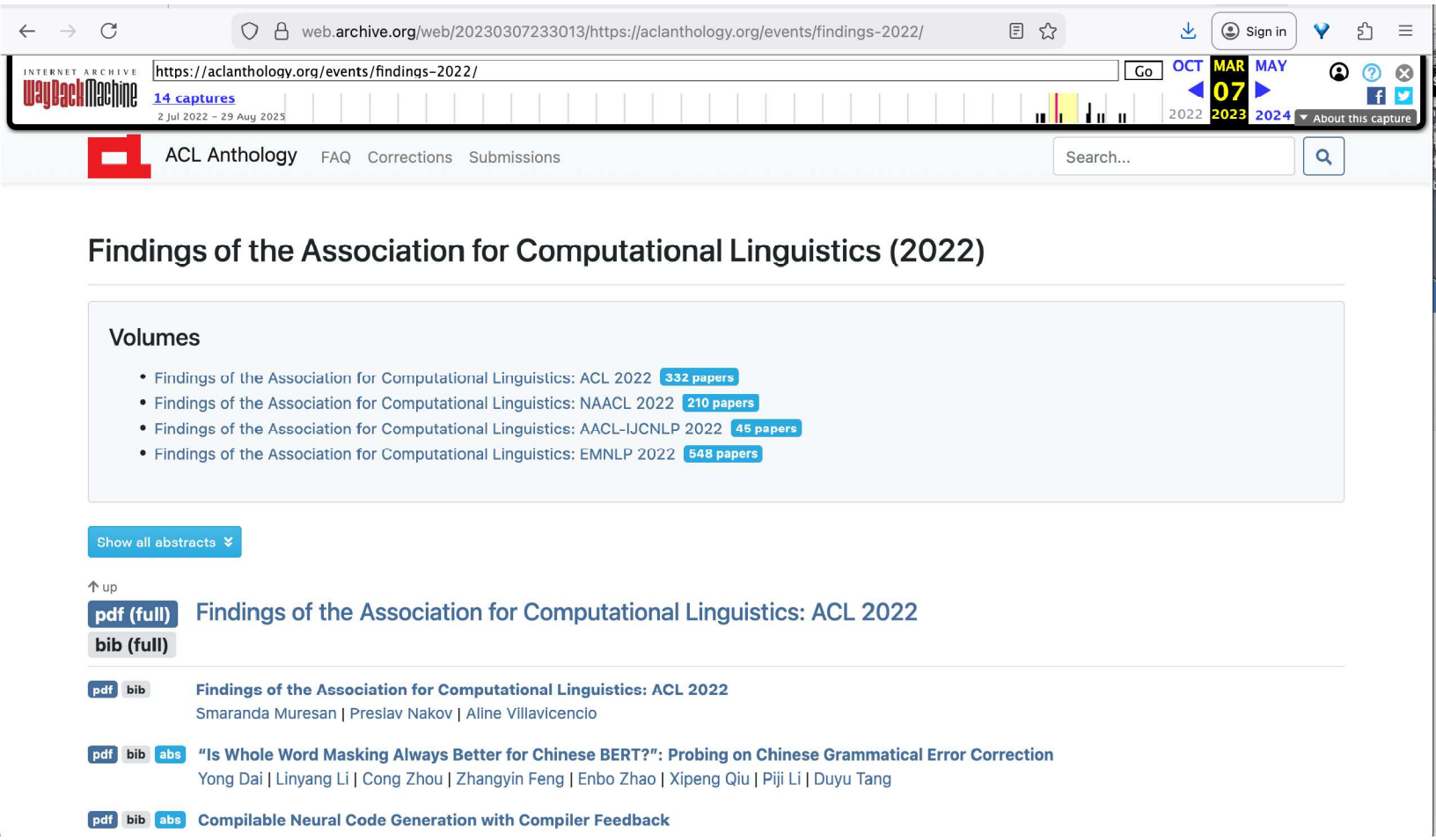
Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Ziyue Wang, Baoxin Wang, Xingyi Duan, Dayong Wu, Shijin Wang, Guoping Hu, and Ting Liu. 2019. IFlyLegal: A Chinese legal system for consultation, law searching, and document analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 97–102, Hong Kong, China. Association for Computational Linguistics.

Moshe Wasserblat. 2021. Sentence transformer fine-tuning (setfit): Outperforming gpt-3 on few-shot text-classification while being 1600 times smaller.

APPENDIX 1011F


web.archive.org/web/20230314133352/https://aclanthology.org/
Sign in
INTERNET ARCHIVE
WaybackMachine
https://aclanthology.org/
Go
670 captures
24 May 2021 - 1 Jan 202
FEB
MAR
14
APR
2022
2023
2024
About this capture
ACL Anthology
FAQ
Corrections
Submissions
Search...
Welcome to the ACL Anthology!
The ACL Anthology currently hosts 83245 papers on the study of computational linguistics and natural language processing.
Subscribe to the mailing list to receive announcements and updates to the Anthology.
Full Anthology as BibTeX (7.16 MB)
...with abstracts (19.65
ACL Events
Venue
2023 – 2020
2019 – 2010
2009 – 2000
1999 – 1990
1989 and older
AACL 22 20
ACL 22 21 20 19 18 17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00 99 98 97 96 95 94 93 92 91 90 89 88 87 86 85 84 83 82 81
ANLP 00 97 94 92 88 83
CL 22 21 20 19 18 17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00 99 98 97 96 95 94 93 92 91 90 89 88 87 86 85 84 83 82 81
CoNLL 22 21 20 19 18 17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00 99 98 97
EACL 21 17 14 12 09 06 03 99 97 95 93 91 89 87 85 83
EMNLP 22 21 20 19 18 17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00 99 98 97 96
Findings 22 21 20
IWSLT 22 21 20 19 18 17 16 15 14 13 12 11 10 09 08 07 06 05 04
NAACL 22 21 19 18 16 15 13 12 10 09 07 06 04 03 01 00
SemEval 22 21 20 19 18 17 16 15 14 13 12 10 07 04 01 98
*SEM 22 21 20 19 18 17 16 15 14 13 12
TACL 22 21 20 19 18 17 16 15 14 13
WMT 22 21 20 19 18 17 16 15 14 13 12 11 10 09 08 07 06
WS 23 22 21 20 19 18 17 16 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00 99 98 97 96 95 94 93 91 90 89 87 85 83 81
SIGs
ANN | BIOMED | DAT | DIAL | EDU | EL | FSM | GEN | HAN | HUM | LEX | MEDIA | MOL | MORPHON | MT | NLL | PARSE | REP | SEM | SEMITIC | SLAV | SLPAT | SLT | TYP | U
Non-ACL Events
Venue
2023 – 2020
2019 – 2010
2009 – 2000
1999 – 1990
1989 and ol


web.archive.org/web/20230307233013/https://aclanthology.org/events/findings-2022/
Sign in
INTERNET ARCHIVE
WayBackMachine
https://aclanthology.org/events/findings-2022/
Go
14 captures
2 Jul 2022 - 29 Aug 2025
OCT
MAR
07
2023
MAY
2022
2024
About this capture
ACL Anthology
FAQ
Corrections
Submissions
Search...
Findings of the Association for Computational Linguistics (2022)
Volumes
Findings of the Association for Computational Linguistics: ACL 2022 332 papers
Findings of the Association for Computational Linguistics: NAACL 2022 210 papers
Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022 45 papers
Findings of the Association for Computational Linguistics: EMNLP 2022 548 papers
Show all abstracts
up
pdf (full)
bib (full)
Findings of the Association for Computational Linguistics: ACL 2022
pdf bib
Findings of the Association for Computational Linguistics: ACL 2022
Smaranda Muresan | Preslav Nakov | Aline Villavicencio
pdf bib abs
"Is Whole Word Masking Always Better for Chinese BERT?": Probing on Chinese Grammatical Error Correction
Yong Dai | Linyang Li | Cong Zhou | Zhangyin Feng | Enbo Zhao | Xipeng Qiu | Piji Li | Duyu Tang
pdf bib abs
Compilable Neural Code Generation with Compiler Feedback

web.archive.org/web/20230307233013/https://aclanthology.org/events/findings-2022/#2022find
Sign in
INTERNET ARCHIVE WayBackMachine
https://aclanthology.org/events/findings-2022/
Go
OCT MAR MAY
14 captures
2 Jul 2022 - 29 Aug 2025
2022 07 2023 2024
About this capture

pdf bib abs **Partitioned Gradient Matching-based Data Subset Selection for Compute-Efficient Robust ASR Training**
Ashish Mittal | Durga Sivasubramanian | Rishabh Iyer | Preethi Jyothi | Ganesh Ramakrishnan

pdf bib abs **Adaptive Graph Convolutional Network for Knowledge Graph Entity Alignment**
Renbo Zhu | Xukun Luo | Meng Ma | Ping Wang

pdf bib abs **Towards Robust NLG Bias Evaluation with Syntactically-diverse Prompts**
Arshiya Aggarwal | Jiao Sun | Nanyun Peng

pdf bib abs **PcMSP: A Dataset for Scientific Action Graphs Extraction from Polycrystalline Materials Synthesis Procedure Text**
Xianjun Yang | Ya Zhuo | Julia Zuo | Xinlu Zhang | Stephen Wilson | Linda Petzold

pdf bib abs **Validity Assessment of Legal Will Statements as Natural Language Inference**
Alice Kwak | Jacob Israelsen | Clayton Morrison | Derek Bambauer | Mihai Surdeanu

pdf bib abs **AdaPrompt: Adaptive Model Training for Prompt-based NLP**
Yulong Chen | Yang Liu | Li Dong | Shuohang Wang | Chenguang Zhu | Michael Zeng | Yue Zhang

pdf bib abs **Code Generation From Flowcharts with Texts: A Benchmark Dataset and An Approach**
Zejie Liu | Xiaoyu Hu | Deyu Zhou | Lin Li | Xu Zhang | Yanzheng Xiang

pdf bib abs **Focus! Relevant and Sufficient Context Selection for News Image Captioning**
Mingyang Zhou | Grace Luo | Anna Rohrbach | Zhou Yu

pdf bib abs **Generative Aspect-Based Sentiment Analysis with Contrastive Learning and Expressive Structure**
Joseph Peper | Lu Wang

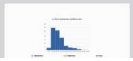
pdf bib abs **Semantic Dependency Parsing with Edge GNNs**
Songlin Yang | Kewei Tu

web.archive.org/web/20230202142404/https://aclanthology.org/2022.findings-emnlp.447.pdf
Sign in
INTERNET ARCHIVE Wayback Machine
https://aclanthology.org/2022.findings-emnlp.447.pdf
Go
2 captures
2 Feb 2023 - 26 Mar 2025
JAN FEB MAR
2022 02 2023 2024
About this capture









Validity Assessment of Legal Will Statements as Natural Language Inference

Alice Saebom Kwak, Jacob O. Israelsen, Clayton T. Morrison, Derek E. Bambauer, Mihai Surdeanu
The University of Arizona, Tucson, Arizona, USA
{alicekwak, jisraelsen, claytonm, derekbambauer, msurdeanu}@arizona.edu

Abstract

This work introduces a natural language inference (NLI) dataset that focuses on the validity of statements in legal wills. This dataset is unique because: (a) each entailment decision requires three inputs: the statement from the will, the law, and the conditions that hold at the time of the testator's death; and (b) the included texts are longer than the ones in current NLI datasets. We trained eight neural NLI models

NLI in general deals with determining whether a premise entails, contradicts, or is neutral to the hypothesis given. Our project made two important changes to this approach to fit legal documents. First, our dataset contains three input types: statements from wills (as hypotheses), laws (as premises), and *conditions*, which are circumstances at the time of probate (e.g., eligibility of the testator or witnesses). This adaptation is necessary

Validity Assessment of Legal Will Statements as Natural Language Inference

Alice Saebom Kwak, Jacob O. Israelsen, Clayton T. Morrison, Derek E. Bambauer, Mihai Surdeanu
The University of Arizona, Tucson, Arizona, USA
{alicekwak, jisraelsen, claytonm, derekbambauer, msurdeanu}@arizona.edu

Abstract

This work introduces a natural language inference (NLI) dataset that focuses on the validity of statements in legal wills. This dataset is unique because: (a) each entailment decision requires three inputs: the statement from the will, the law, and the conditions that hold at the time of the testator's death; and (b) the included texts are longer than the ones in current NLI datasets. We trained eight neural NLI models in this dataset. All the models achieve more than 80% macro F1 and accuracy, which indicates that neural approaches can handle this task reasonably well. However, group accuracy, a stricter evaluation measure that is calculated with a group of positive and negative examples generated from the same statement as a unit, is in mid 80s at best, which suggests that the models' understanding of the task remains superficial. Further ablative analyses and explanation experiments indicate that all three text segments are used for prediction, but some decisions rely on semantically irrelevant tokens. This indicates that overfitting on these longer texts likely happens, and that additional research is required for this task to be solved.

1 Introduction

Natural language inference (NLI) in the legal domain has not been widely investigated, despite its importance and potential. One such important NLI application is validity assessment of legal documents such as wills. Legal procedures for creating and executing wills are evolving rapidly. Processing a will via probate is a costly, time-consuming process that can be exacerbated by errors or by challenges to validity. These problems will likely increase as people increasingly employ electronic wills. Developing natural language techniques that can determine a will's validity at creation and execution can increase validity, conserve legal resources, and effectuate the author's intent. To this end, this work explores how NLI models can be employed to evaluate the validity of will statements.

NLI in general deals with determining whether a premise entails, contradicts, or is neutral to the hypothesis given. Our project made two important changes to this approach to fit legal documents. First, our dataset contains three input types: statements from wills (as hypotheses), laws (as premises), and *conditions*, which are circumstances at the time of probate (e.g., eligibility of the testator or witnesses). This adaptation is necessary as will statements' validity cannot be evaluated without considering both circumstances at the time of probate ("conditions") and relevant legal rules ("laws"). Also, we switched the labels from entailment, contradict, neutral to support, refute, unrelated to better represent the relation between will statements and laws.
The major contributions of this work are:

- We create an open-access annotated dataset with 1,014 data points, generated from 23 publicly available wills. In addition to the three-tuple setting unique to the legal domain, this dataset also contains texts considerably longer than in other open-domain NLI datasets, such as SNLI (Bowman et al., 2015). The average length of our texts is 269 tokens, while the average lengths of premises and hypotheses in SNLI are 8 and 14 tokens, respectively.

- We demonstrate that validity assessment of legal will statements can be handled reasonably well with state-of-the-art NLI models when trained on our dataset. However, low scores in group accuracy, which is a stricter evaluation measure calculated with a group of positive and negative examples generated from the same statement as a unit, indicate that work remains before NLI models fully understand legal language.

- We explain how the trained models utilize our dataset via ablation tests that remove various

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6076 - 6085
December 7-11, 2022 ©2022 Association for Computational Linguistics
Exhibit 1026 Page 161

pieces of information at prediction time (laws, conditions, or both), and through post-hoc explainability analyses using LIME (Ribeiro et al., 2016). Our analyses indicate that in most cases the NLI classifiers use meaningful phrases from all three pieces of text, indicating the models do capture useful information. However, in some situations, the models use features that are not intuitive for humans, indicating that the task is incomplete.

2 Related Work

Natural language inference (NLI), also known as Recognizing Textual Entailment (RTE), determines entailment relations between a pair of sentences: a premise and a hypothesis. Their relationship is either: a) entailment, if a premise entails a hypothesis; b) contradiction, if a premise contradicts a hypothesis; or c) neutral, if a premise neither entails nor contradicts a hypothesis. NLI has been a key framework in natural language processing since Dagan et al. (2006) proposed the RTE challenge.

Numerous datasets exist for general NLI tasks. For example, the Stanford Natural Language Inference (SNLI) Corpus (Bowman et al., 2015) dataset contains about 570,000 pairs of sentences (premises and hypotheses) generated from Flickr image captions. More recently, datasets for domain-specific NLI tasks are being introduced. SciNLI (Sadat and Caragea, 2022) is drawn from scientific texts. It contains 107,412 sentence pairs extracted from papers in natural language processing and computational linguistics.

There have been a few NLI-related resources for the legal domain, including IFlyLegal (Wang et al., 2019), StAtutory Reasoning Assessment (SARA) dataset (Holzenberger et al., 2020), Graph-based Causal Inference (GCI) framework (Liu et al., 2021), and ContractNLI (Koreeda and Manning, 2021). One IFlyLegal module was developed for natural language article inference; it provides relevant legal articles when asked legal questions. SARA contains rules extracted from the US Internal Revenue Code and natural language questions. GCI generates causal graphs based on fact descriptions. Though the framework was not specifically designed for law, Liu et al. (2021) demonstrated that it can be utilized for legal text analysis. ContractNLI is a dataset consisting of 607 legal contracts designed to handle document-level natural language inference.

Our work differs from these existing NLI works in two ways: (a) it contains three types of input information; and (b) it operates in a pragmatic middle space with respect to text length: larger than datasets with sentence-level texts, which are insufficient to capture legal details, but shorter than document-level inference datasets.

3 Dataset

Our dataset includes three types of entries (will statements, applicable law, and facts representing external state) rather than the usual two types. Validity often depends upon such external circumstances. Thus, a will statement can be legally valid under some facts, but not others. Hence, a will can change from valid to invalid or vice versa; wills must, at minimum, be evaluated when executed (drafted and signed) and when probated (where a court determines whether to implement its provisions). For example, if a will contains a statement appointing a specific person as executor, that person's eligibility must be verified. This requires information about both the law and the person's circumstances at probate time. For example, according to Tennessee Code section 40-20-115, any person who has been imprisoned cannot be an executor. Thus, to assess eligibility of a Tennessee executor in Tennessee, one must know whether they have been in prison at probate time.

3.1 Data Collection

Our data was collected from the U.S. Wills and Probates datasets in Ancestry, which contains documents in the public domain.[1] We chose 23 wills based on three criteria: 1) whether the wills were handwritten; 2) the wills' execution and probation dates; and 3) where the wills were executed and probated. We excluded handwritten wills due to the difficulty of OCR text recognition. Execution and probation dates can affect interpretation of wills; we excluded wills executed before 1970 and probated before 2000. Lastly, we only collected wills from Tennessee, because including wills from multiple states would require analyzing each state's laws governing wills. Tennessee had the greatest number of wills meeting our criteria.

All collected wills were anonymized. Personal information was replaced with special tokens denoting the type of information (such as [Person-

[1]Court documents are in the public domain in the US.

Exhibit 1026 Page 162

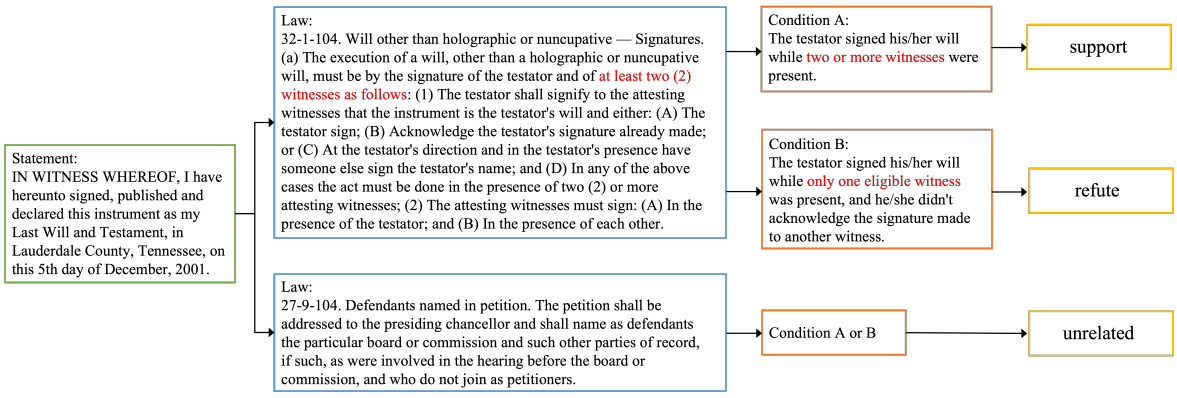


Figure 1: A demonstration of the annotation procedure. Given a law and condition, each statement is evaluated for validity. *Support*: it is supported by the law and condition provided; *Refute*: refuted by the law and condition. *Unrelated*: not relevant to the law. When a statement and a law are unrelated to each other, a condition is not required for classification. However, it is still assigned one of the conditions from either support or refute case. The goal is to prevent NLI models from depending on contextual difference (i.e., "texts without conditions are all unrelated cases") rather than features when making predictions.

n], [Address-n], [Number-n], where n identified each person or object in the will. (Suntwal et al. (2019) suggested this anonymization method). Anonymization was performed manually to prevent including personal information in the dataset.

3.2 Annotation

Each will statement was annotated as *support*, *refute*, or *unrelated* based on a given condition and law. (Our annotators set hypothetical conditions when necessary to categorize statements.)

Importantly, each statement was annotated multiple times. A statement was annotated once as support, once as refute, and thrice as unrelated.[2] This ensured that every statement was labeled with all three classes in the same ratio (support:refute:unrelated=1:1:3). Also, annotating statements three times more with "unrelated" than support and refute enabled including a greater range of laws in the dataset. Formally, our annotation procedure included the following steps:

(1) preprocessing: extracting texts from collected wills using OCR[3], and copying statements into the dataset;

(2) for each statement, identifying the Tennessee legal provision that supported, refuted, or was unrelated to the statement's validity;

(3) adding a condition specifying external circumstances relevant to validity;

(4) repeating these steps to generate five annotations per statement;

(5) anonymizing all statements by replacing personal information with special tokens.

Figure 1 demonstrates the annotation procedure with examples.

Two annotators participated in the task. One is a law student, and annotation was supervised by author Bambauer, a law professor. The other annotator does not have legal training, but ongoing discussions and reviews ensured the uniform quality. The annotators contributed equally. Annotators worked individually but shared annotation guidelines. After the annotation was complete, cross-annotation was conducted on 200 randomly selected items (100 items from each annotator's set; each annotator worked blinded on items drawn from the other annotator's set) to check inter-annotator agreement. We calculated Kappa agreement score based on the cross-annotation result, and the score is 0.91 (rounded to hundredth).

After the annotation was complete, the dataset was split into training, development, and test sets. There are 1014 data points in our dataset, split 50:25:25% (train: 504, development: 255, test: 255). When splitting, a group containing annotations from a single statement (one support, one refute, and three unrelated) was treated as a single

[2] A few exceptions occurred when:
a) it was impossible to create a case where the statement was refuted. In such cases, statements were not annotated as "refute."
b) a statement had more than one relevant legal rule. Such statements were annotated twice as support, twice as refute, and once as unrelated.

[3] The collected wills were scanned documents, so OCR was necessary to extract machine-readable text.

Exhibit 1026 Page 163

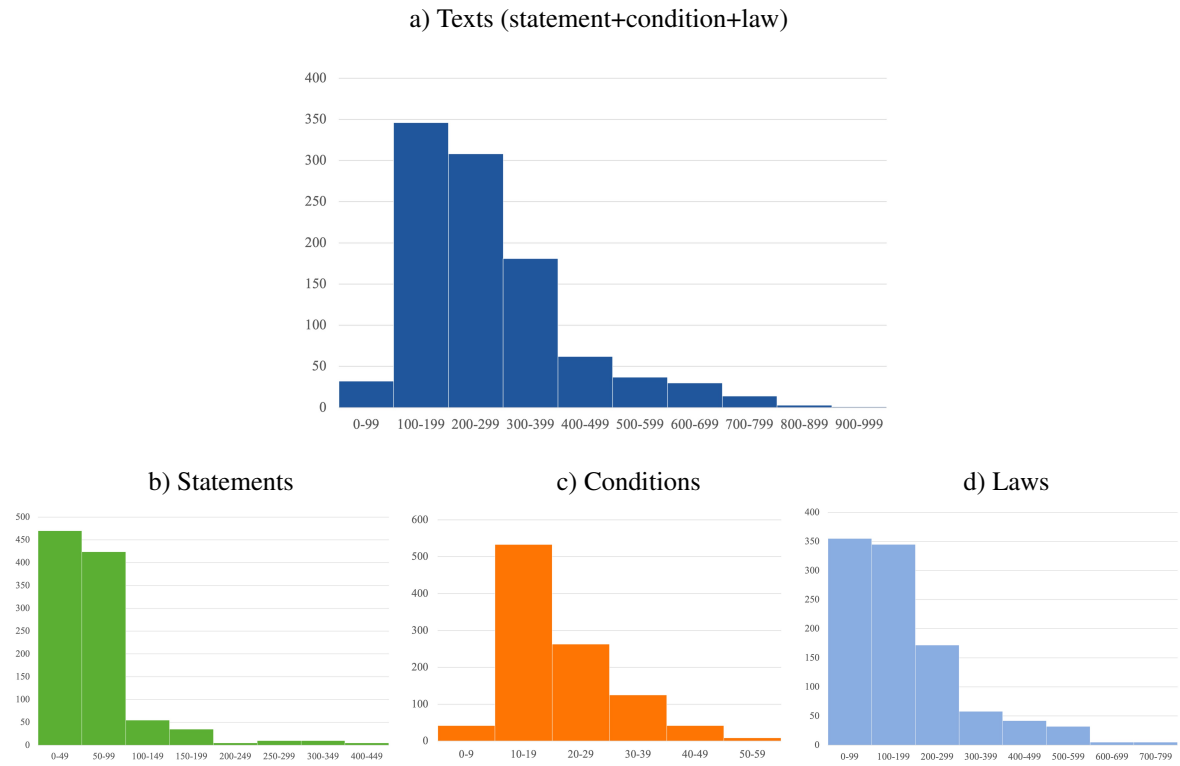


Figure 2: Histograms demonstrating the length distribution of a) full texts (i.e., statement+condition+law), b) statements, c) conditions, and d) laws in terms of token counts. Token counts are plotted on the X-axis; the Y-axis indicates the number of texts/statements/conditions/laws for each bin.

Datasets	**Texts**	**Tokens**
Train	504	135218
Development	255	69643
Test	255	68042
Total	**1014**	**272903**

Table 1: The number of texts and tokens contained in each dataset and in total

unit. Each annotator's work was equally represented in all three datasets. All 23 wills appear in more than one of the train, dev, and test partitions. However, given the nature of the will statements from the wills and how they were annotated, data leakage is unlikely. Will statements are independent and generally do not convey information about other statements from the same wills. Summary statistics are in Table 1 and Figure 2.

3.3 Characteristics of the Dataset

Our dataset has two characteristics distinguishing it from other datasets: a) texts are composed of three input types, and b) texts contain a large number of tokens.

First, our task requires a third input type: a condition, since validity often depends upon conditions.

Second, our texts tend to contain a large number of tokens. 44% of our texts contain 200 or more tokens. The average token number was 269.14; 79 texts contain more than 512 tokens (the threshold for most NLI models). Popular NLI datasets are shorter.

3.4 Implementation

We trained multiple NLI models with our dataset to assess their performance: four transformer models and four sentence-transformer models. The transformer models include bert-base-uncased (Devlin et al., 2018), distilbert-base-uncased (Sanh et al., 2019a), roberta-large-mnli (Liu et al., 2019), and longformer-base-4096 (Beltagy et al., 2020). Bert-base-uncased and dilstilbert-base-uncased were trained to set baselines on the task. In addition to the baseline models, we used roberta-large-mnli and longformer-base-4096. Among sentence-transformer models (Reimers and Gurevych, 2019), four pretrained models with top average performances (based on the Model Overview page[4]) were chosen: all-mpnet-base-v2[5], multi-qa-mpnet-base-

[4] https://www.sbert.net

[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Exhibit 1026 Page 164

dot-v1[6], all-distilroberta-v1 (Sanh et al., 2019b)[7], and all-MiniLM-L12-v2 (Wang et al., 2020)[8].

To distinguish between statements, laws, and conditions in the concatenated texts, we prefixed each with a special token: `[STATE]`, `[LAW]`, and `[COND]`.

Both transformer models and sentence-transformer models were trained on PyTorch 1.11.0 with Cuda 11.3. All transformer models were trained using the Trainer class provided by HuggingFace. Learning rates and training epochs were tuned on the development partition. Sentence-transformer models were trained with Sentence Transformer Fine-Tuning (SetFit) proposed by Wasserblat (2021). It was slightly adapted to fit the multi-class classification task, but its fundamentals remain intact. SetFit utilizes sentence pairs generated within the same class as training data. It fits the model with the data to minimize the Softmax Loss. Once the model is fitted, it is used to encode the training and development (or test, when it is a testing phase) datasets. The encoded data is used to fit the Logistic Regression classifier (with 'liblinear' solver) which, finally, makes predictions.

All models except for longformer-base-4096 were trained and evaluated on the truncated datasets due to the models' token number limitations. The models can only process texts with 512 or fewer tokens without truncation (all-MiniLM-L12-v2: 256 tokens; all-mpnet-base-v2: 384 tokens; other models, except longformer-base-4096: 512 tokens). When truncating datasets, the ratio of average token numbers among three input types (statements, laws, and conditions) was considered to prevent losing excessive information from a single type of input.

4 Results and Analysis

4.1 Evaluation Measures

We report results using standard accuracy, precision, recall, and F1 scores. To ensure all labels are well represented, precision, recall, and F1 scores were computed with Macro average. Additionally, as suggested by Elazar et al. (2021), we introduce a new measure called group accuracy (GA). GA is calculated with a group of positive and negative examples from the same statement (rather than each text) as a unit. If a group has one or more incorrect predictions, the group is incorrect. If a model correctly understands a will statement, it should perform equally well on all examples from the same statement.

4.2 Results from Trained Models

Table 2 shows the models' performances on our test partition. Overall, the models demonstrated good performances. Each achieved more than 80% in all metrics but group accuracy. Roberta-large-mnli showed the best performance. It achieved over 96% in all metrics but group accuracy (84.31%), suggesting it can handle the task reasonably well. However, the large difference between accuracy and group accuracy suggests that the models' understanding of the task is superficial.

Accuracy for the *unrelated* label was higher than for *support* and *refute* in all models except all-mpnet-base-v2 and multi-qa-mpnet-base-dot-v1. This higher accuracy for *unrelated* label may be partially attributable to the gap between accuracy and group accuracy, as it would inflate overall accuracy. However, since the gap between accuracy and group accuracy is not significantly smaller for all-mpnet-base-v2 and multi-qa-mpnet-base-dot-v1 (i.e., the models where accuracy for *unrelated* was not higher than the other labels), it is more likely that the gap originated from the strictness of group accuracy and the models' superficial understanding of will statements.

Further, roberta-large-mnli (trained with truncated inputs) showed better performance than longformer-base-4096 (trained with full length inputs). This suggests that models with token number limitations can still perform well with long inputs when truncated properly.

4.3 Experiment with Different Input Lengths

To investigate whether the models' performance was affected by input length and/or input truncation, we varied input lengths. Inputs were classified into three categories by length: a) short (equal or less than 192 or 256 tokens); b) regular (larger than 192 or 256 tokens and equal or less than 384 or 512); and c) long (larger than 384 or 512)[9]. The

[6] https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1

[7] https://huggingface.co/sentence-transformers/all-distilroberta-v1

[8] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

[9] For all-mpnet-base-v2 model, 192 and 384 were used as dividing points instead of 256 and 512, as its maximum token length is 384.

Exhibit 1026 Page 165

Model	Precision	Recall	F1	Accuracy	GA
Transformers models					
bert-base-uncased	81.47	82.20	81.81	87.06	49.02
distilbert-base-uncased	82.37	82.59	82.48	87.06	50.98
longformer-base-4096	94.09	93.69	93.85	94.90	74.51
roberta-large-mnli	**96.67**	**96.25**	**96.42**	**96.86**	**84.31**
Sentence-Transformers models					
all-MiniLM-L12-v2	81.50	84.54	82.64	85.49	50.98
all-disilroberta-v1	83.43	85.99	84.22	87.06	49.02
multi-qa-mpnet-base-dot-v1	91.48	94.13	92.65	93.73	74.51
all-mpnet-base-v2	92.93	95.05	93.82	94.90	76.47

Table 2: Results of all classifiers trained on our dataset using five measures: precision, recall, F1, accuracy, and group accuracy (GA in table, see Section 4.1). Precision, recall, and F1 scores were computed with Macro average. Group accuracy is calculated with positive and negative examples generated from the same statement as a unit. If a group contains one or more incorrect predictions, the group is considered incorrect.

Model	Precision	Recall	F1	Accuracy
all-mpnet-base-v2 (sample n = 52)				
token <= 192	93.75	93.15	92.91	94.23
192 < token <= 384	90.48	89.10	89.63	94.23
384 < token	86.35	89.49	87.08	88.46
longformer-base-4096 (sample n = 20)				
token <= 256	95.24	88.89	90.77	95.00
256 < token <= 512	80.56	71.67	70.93	80.00
512 < token	73.81	66.67	66.93	75.00
roberta-large-mnli (sample n = 20)				
token <= 256	100	100	100	100
256 < token <= 512	94.87	85.00	88.76	90.00
512 < token	94.44	93.33	93.27	95.00

Table 3: Results with different input lengths. Models generally performed worse when token numbers in inputs were larger. Roberta-large-mnli performed better than longformer-base-4096 in predicting long inputs (token > 512) despite truncation. Roberta-large-mnli also showed better performance on long inputs than regular inputs, indicating input truncation did not worsen the model's performance.

number of inputs in each category varied. To control the impact of varying sample size, categories with larger sample sizes were reduced by random sampling to match the smallest sample size. For this experiment, we used the three models with best performance on the full dataset.

Table 3 shows the results. Models generally performed worse with larger input token numbers. All three models performed best with inputs with smaller token numbers. One interesting finding is that roberta-large-mnli performed better than longformer-base-4096 in predicting inputs with large token numbers (n > 512) despite truncation. Roberta-large-mnli showed better performance on long inputs than regular inputs, indicating truncation did not negatively affect the model's performance. This finding aligns with the observation from the overall results that models with token number limitation can still show good performances on long inputs when properly truncated. However, results differed with a smaller token number limitation. All-mpnet-base-v2 performed worst on long inputs (where truncation occurred), indicating truncation negatively impacted performance.

4.4 Results from Ablation Experiment

To determine whether the models correctly rely on all three types of texts (statements, conditions, and laws), we conducted an ablation experiment, training the models with datasets lacking one (law or condition) or two types (law+condition) of inputs. Poliak et al. (2018) suggested a similar experiment

Exhibit 1026 Page 166

Model	Precision	Recall	F1	Accuracy	GA
Statements and laws					
all-mpnet-base-v2	75.64	76.47	75.98	81.96	27.45
longformer-base-4096	75.72	75.98	75.79	82.75	37.25
roberta-large-mnli	77.93	77.51	77.55	84.31	39.22
Statements and conditions					
all-mpnet-base-v2	46.67	57.13	44.80	44.31	0.00
longformer-base-4096	58.00	49.44	51.30	58.82	0.00
roberta-large-mnli	37.63	41.72	38.08	56.86	0.00
Statements only					
all-mpnet-base-v2	36.76	35.94	34.04	36.47	0.00
longformer-base-4096	18.17	33.33	23.52	54.51	0.00
roberta-large-mnli	18.17	33.33	23.52	54.51	0.00

Table 4: Results from the ablation experiment where three models were trained with datasets lacking one (law or condition) or two types (law+condition) of inputs. Results demonstrate that the models' performances deteriorate if one or more input type(s) is omitted. Thus, both types of inputs affect the models' performances, though laws have larger impact than conditions.

to this one to set hypothesis-only baselines. We used the three models with the best performances on the full dataset for this experiment.

Table 4 shows the results. Performance deteriorated if any input type was omitted. Among the models trained on partial data, those trained with statements and laws performed better, with F1 scores over 75%. However, group accuracy dropped to the 30s or even 20%s, indicating actual understanding decreased considerably. Models trained with statements and conditions performed significantly worse. Results ranged between 38% and 52% for F1, and group accuracy dropped to 0. Models trained only with statements performed worst, with F1 scores dropping to 35% or lower.

This degradation from partial data shows including both conditions and laws positively affects the models' performance. Models use all three types of inputs when making predictions[10]. Laws have more impact on performance than conditions, since deterioration in results without laws was more significant than from models trained without conditions. This is realistic, since laws set the parameters within which conditions operate to make a given provision valid or invalid.

4.5 Understanding Results with LIME

To clarify model behavior, we used Local Interpretable Model-Agnostic Explanations (LIME) to explain our classifiers' predictions (Ribeiro et al., 2016). We implemented LIME on predictions made by the best performing model (roberta-large-mnli, fine-tuned on our data). The LIME results revealed that the model tends to correctly rely on all three texts (statements, laws, conditions) for a prediction, and uses meaningful features in many cases. However, sometimes the model utilized features that are not intuitive for humans.

4.5.1 A Correct Example

Figure 3 shows a case where the model made a correct prediction based on features sensible to humans. Figure 3's text is a statement saying two or more witnesses witnessed the testator signing the will and signed in each other's presence. It includes a condition stating that one of the two witnesses was ineligible to serve under Tennessee law, and the law specifying witness eligibility in Tennessee. The given condition and the law invalidate the will statement, rendering it as *refute*. The model correctly predicted the outcome based on tokens such as *ineligible*, *witnesses*, *One*, *testament*, and *2*. The top two tokens with greatest impact on the model's prediction were *ineligible* and *witnesses*, with importance score of 0.60 and 0.14, respectively. It is

[10] A reviewer suggested an experiment where laws and conditions are packed into premises and treat the task as normal NLI. A pilot experiment with the best performing model (roberta-large-mnli) found a slight deterioration in performance when laws and conditions were combined as premises (precision: 93.09, recall: 93.89, F1: 93.41, accuracy: 94.12, GA: 74.51; see Table 2 for comparison with main result). The biggest difference came from GA score (74.51 vs. 84.31). This lowered GA score implies that the model's understanding worsened when conditions and laws were combined into premises. Given this preliminary result, it is likely that three-way distinction in input types positively contributes to the model's performance. Nevertheless, as the performance difference is not significantly large, further investigation is needed.

Exhibit 1026 Page 167

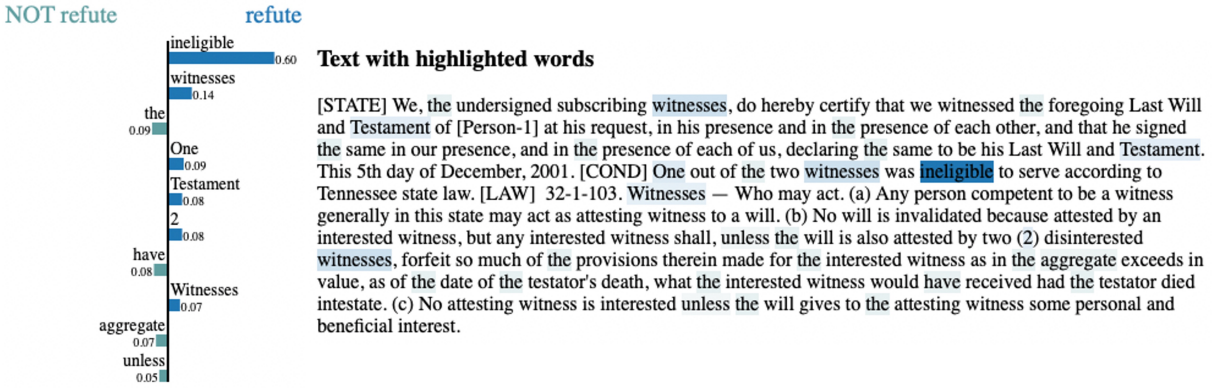


Figure 3: Examples of LIME explanations showing a case where the model (roberta-large-mnli) makes correct predictions based on features sensible to humans.

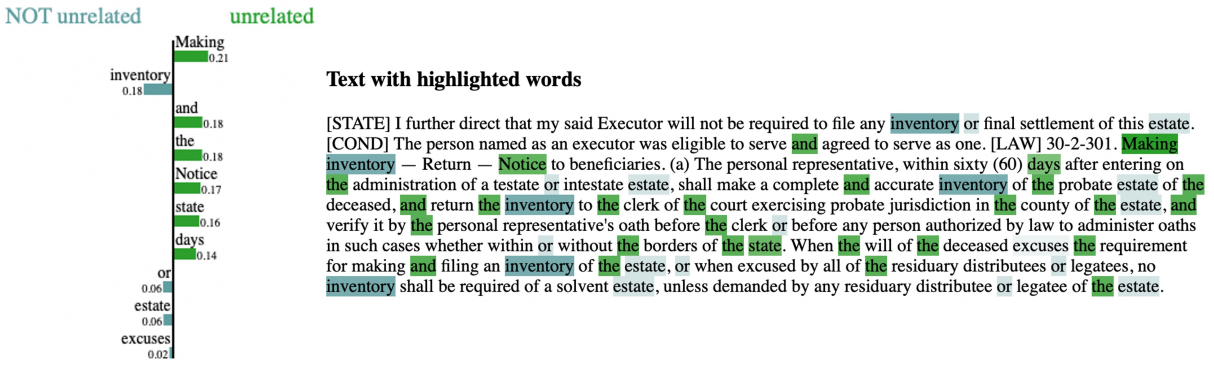


Figure 4: An example of a LIME explanation showing a case where the model (roberta-large-mnli) makes an incorrect prediction based on the tokens such as *Making*, *and*, and *the* which bear little semantic relevancy to the gist of the text.

understandable that these tokens have high scores, as they are the keywords ("**One** out of two (**2**) **witnesses** is **ineligible**") which provide grounds for revoking a will statement.

4.5.2 An Incorrect Example

Figure 4 shows a case where the model made an incorrect prediction based on irrelevant tokens. The intent of the will statement is to excuse the Executor from filing inventory. This is not contradicted by the condition and law, requiring the model to classify the text as *support*. Nevertheless, the model incorrectly classified it as *unrelated*. The reason for the incorrect prediction is that the model (probably) relied on irrelevant tokens such as *Making*, *and*, and *the*, which bear little semantic relevancy to the text, versus more relevant tokens such as *inventories* and *excuses* when making the prediction. This is likely due to overfitting induced by longer texts. This LIME explanation shows that even the best performing model still has room for improvement.

5 Future Work

This work can be expanded in several directions. First, it can be extended to cover multiple states by adapting the models or adding more data. Future work can investigate if models trained on a single state's data can be adapted to evaluate data from other states. Also, wills from multiple states could be added to the dataset. We expect the augmented dataset would enhance the models' performance on evaluating wills from other states.

This work can be expanded to other legal domains. Models trained on our data can be adapted to similar tasks in other legal domains.

Lastly, this work can be extended by investigating novel transformer models. Given the uniqueness of our dataset with regard to the number of inputs and text length, it is likely that further experiment and modification is needed to handle such characteristics.

Exhibit 1026 Page 168

6 Conclusion

This work presented an annotated dataset for natural language inference (NLI) in the legal domain, consisting of 1,014 data points generated from 23 publicly available wills. The dataset is novel for two reasons: it included texts with three input types (statement, law, and condition) rather than two (premise and hypothesis) in the traditional NLI, and included texts are longer than in general NLI datasets. The NLI models trained on our dataset showed reasonable performance in assessing the validity of will statements. Ablative experiments demonstrated that the models' performances worsen if any input type (condition, law, or both) is omitted. This suggests that the models utilize all three input types. The LIME implementation reveals that even the best performing model makes errors in some cases by using semantically irrelevant tokens. Our open-access dataset is publicly available at: https://github.com/ml4ai/nli4wills-corpus

Limitations

Our dataset consists of a relatively small number of data points (1,014 texts). Annotating will statements with relevant laws and conditions is a highly demanding, time intensive task. A larger size of dataset would likely further improve the models' performances.

Our dataset only includes wills executed and probated in Tennessee, with execution after 1970 and probate after 2000. Due to these restrictions, our framework might produce incorrect results on inputs from wills from different settings. Supplementing the dataset with wills from more diverse settings would address this limitation. Even though the scope is limited to a single state (Tennessee), our study demonstrates that transformer models trained on the dataset can evaluate the validity of statements from wills with reasonable accuracy. Future additions to our dataset will be available at the same URL: https://github.com/ml4ai/nli4wills-corpus

This work does not involve humans in the loop. Considering how crucial accuracy is for the task (i.e., legal validity evaluation), the work would have benefited much from involving domain experts. Even though our study discovered that the state-of-the-art transformer models can show good performances (over 85% accuracy in all 8 models) without human interaction, it was also found that the models' understanding on the task is rather superficial (GAs ranging from 49-85% in Table 2). Including humans-in-the-loop could be a solution for enhancing the models' understanding on the task.

Ethics Statement

We collected legal wills from Ancestry as a part of our dataset creation process. The wills probated in court are in the public domain in the US, and we did not violate Ancestry's Terms and Conditions when collecting wills. We also anonymized the wills by replacing any personal identifiable information contained in the documents with special tokens.

Our datasets and codes are released to public. We believe our released datasets and codes would contribute to society by promoting further NLI endeavors in the legal domain. The resources could potentially assist with people reviewing wills, but they should not be considered as legal advice. To avoid any confusion, we placed a disclaimer that the users must not rely on any information provided from our resources when making legal decisions and should instead consult with an attorney.

Acknowledgements

We thank the reviewers for their thoughtful comments and suggestions. This work was partially supported by the National Science Foundation (NSF) under grant 2217215, and by University of Arizona's Provost Investment Fund. Mihai Surdeanu and Clayton Morrison declare a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification,*

and Recognising Tectual Entailment, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

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.

Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. 2021. Back to square one: Artifact detection, training and commonsense disentanglement in the Winograd schema. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10486–10500, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering. *CoRR*, abs/2005.05257.

Yuta Koreeda and Christopher D. Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. *CoRR*, abs/2110.01799.

Xiao Liu, Da Yin, Yansong Feng, Yuting Wu, and Dongyan Zhao. 2021. Everything has a cause: Leveraging causal inference in legal text analysis. *CoRR*, abs/2104.09420.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Mobashir Sadat and Cornelia Caragea. 2022. SciNLI: A corpus for natural language inference on scientific text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7399–7409, Dublin, Ireland. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019a. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019b. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Sandeep Suntwal, Mithun Paul, Rebecca Sharp, and Mihai Surdeanu. 2019. On the importance of delexicalization for fact verification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3413–3418, Hong Kong, China. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Ziyue Wang, Baoxin Wang, Xingyi Duan, Dayong Wu, Shijin Wang, Guoping Hu, and Ting Liu. 2019. IFlyLegal: A Chinese legal system for consultation, law searching, and document analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 97–102, Hong Kong, China. Association for Computational Linguistics.

Moshe Wasserblat. 2021. Sentence transformer fine-tuning (setfit): Outperforming gpt-3 on few-shot text-classification while being 1600 times smaller.

APPENDIX 1011G

papers.ssrn.com/sol3/papers.cfm?abstract_id=4279728

Sign in

SSRN

Product & Services | Subscribe | Submit a paper | Browse | Blog | More | Create Account | Sign In

THE UNIVERSITY OF ARIZONA James E. Rogers College of Law

Download This Paper | Open PDF in Browser

Add Paper to My Library

Share:

Validity Assessment of Legal Will Statements as Natural Language Inference

Findings of the Association for Computational Linguistics: EMNLP 2022 (forthcoming 2022)

Arizona Legal Studies Discussion Paper No. 22-21

11 Pages • Posted: 17 Nov 2022 • Last revised: 21 Nov 2022

Alice Kwak
University of Arizona - Department of Linguistics

Jacob Israelsen
University of Arizona - James E. Rogers College of Law

Clayton Morrison
University of Arizona - School of Information

Derek E. Bambauer
University of Florida Levin College of Law

Do you have a job opening that you would like to promote on SSRN?

Place Job Opening

Paper statistics

DOWNLOADS	ABSTRACT VIEWS	RANK
92	1,308	676,193

PlumX Metrics

Derek E. Bambauer
University of Florida Levin College of Law

Mihai Surdeanu
University of Arizona - Department of Computer Science

Date Written: November 15, 2022

Abstract

This work introduces a natural language inference (NLI) dataset that focuses on the validity of statements in legal wills. This dataset is unique because: (a) each entailment decision requires three inputs: the statement from the will, the law, and the conditions that hold at the time of the testator's death; and (b) the included texts are longer than the ones in current NLI datasets. We trained eight neural NLI models in this dataset. All the models achieve more than 80% macro F1 and accuracy, which indicates that neural approaches can handle this task reasonably well. However, group accuracy, a stricter evaluation measure that is calculated with a group of positive and negative examples generated from the same statement as a unit, is in mid 80s at best, which suggests that the models' understanding of the task remains superficial. Further ablative analyses and explanation experiments indicate that all three text segments are used for prediction, but some decisions rely on semantically irrelevant tokens. This indicates that overfitting on these longer texts likely happens, and that additional research is required for this task to be solved.

Keywords: natural language, natural language inference, NLI, machine learning, smart contracts, blockchain, AI, artificial intelligence, algorithm, automation, texts, legal wills

Suggested Citation:

Kwak, Alice and Israelsen, Jacob and Morrison, Clayton and Bambauer, Derek E. and Surdeanu, Mihai, Validity Assessment of Legal Will Statements as Natural Language Inference (November 15, 2022). Findings of the Association for Computational Linguistics: EMNLP 2022 (forthcoming 2022), Arizona Legal Studies Discussion Paper No. 22-21, Available at SSRN: https://ssrn.com/abstract=4279728

Related eJournals

University of Arizona James E. Rogers College of Law Legal Studies Research Paper Series

Follow

Wills, Trusts, & Estates Law eJournal

Follow

View more >

Show Contact Information >

Download This Paper | Open PDF in Browser

0 References

0 Citations

Submit a Paper >

Section 508 Text Only Pages

SSRN Quick Links

- SSRN Solutions
- Research Paper Series
- Conference Papers
- Partners in Publishing
- Jobs & Announcements
- Special Topic Hubs

SSRN Rankings

- Top Papers
- Top Authors
- Top Organizations

About SSRN

- Network Directors
- Announcements
- Contact us
- FAQs



THE UNIVERSITY OF ARIZONA
James E. Rogers College of Law

Arizona Legal Studies

Discussion Paper No. 22-21

Validity Assessment of Legal Will Statements as Natural Language Inference

Alice Saebom Kwak
The University of Arizona
Department of Linguistics

Jacob O. Israelsen
The University of Arizona
James E. Rogers College of Law

Clayton T. Morrison
The University of Arizona
School of Information

Derek E. Bambauer
The University of Arizona
James E. Rogers College of Law

Mihai Surdeanu
The University of Arizona
Department of Computer Science

November 2022

Electronic copy available at: https://ssrn.com/abstract=4279728

Validity Assessment of Legal Will Statements as Natural Language Inference

Alice Saebom Kwak[1], **Jacob O. Israelsen**[2], **Clayton T. Morrison**[3], **Derek E. Bambauer**[2], **Mihai Surdeanu**[4]

[1]Department of Linguistics, The University of Arizona
[2]James E. Rogers College of Law, The University of Arizona
[3]School of Information, The University of Arizona
[4]Department of Computer Science, The University of Arizona
{alicekwak, jisraelsen, claytonm, derekbambauer, msurdeanu}@arizona.edu

Abstract

This work introduces a natural language inference (NLI) dataset that focuses on the validity of statements in legal wills. This dataset is unique because: (a) each entailment decision requires three inputs: the statement from the will, the law, and the conditions that hold at the time of the testator's death; and (b) the included texts are longer than the ones in current NLI datasets. We trained eight neural NLI models in this dataset. All the models achieve more than 80% macro F1 and accuracy, which indicates that neural approaches can handle this task reasonably well. However, group accuracy, a stricter evaluation measure that is calculated with a group of positive and negative examples generated from the same statement as a unit, is in mid 80s at best, which suggests that the models' understanding of the task remains superficial. Further ablative analyses and explanation experiments indicate that all three text segments are used for prediction, but some decisions rely on semantically irrelevant tokens. This indicates that overfitting on these longer texts likely happens, and that additional research is required for this task to be solved.

1 Introduction

Natural language inference (NLI) in the legal domain has not been widely investigated, despite its importance and potential. One such important NLI application is validity assessment of legal documents such as wills. Legal procedures for creating and executing wills are evolving rapidly. Processing a will via probate is a costly, time-consuming process that can be exacerbated by errors or by challenges to validity. These problems will likely increase as people increasingly employ electronic wills. Developing natural language techniques that can determine a will's validity at creation and execution can increase validity, conserve legal resources, and effectuate the author's intent. To this end, this work explores how NLI models can be employed to evaluate the validity of will statements.

NLI in general deals with determining whether a premise entails, contradicts, or is neutral to the hypothesis given. Our project made two important changes to this approach to fit legal documents. First, our dataset contains three input types: statements from wills (as hypotheses), laws (as premises), and *conditions*, which are circumstances at the time of probate (e.g., eligibility of the testator or witnesses). This adaptation is necessary as will statements' validity cannot be evaluated without considering both circumstances at the time of probate ("conditions") and relevant legal rules ("laws"). Also, we switched the labels from entailment, contradict, neutral to support, refute, unrelated to better represent the relation between will statements and laws.

The major contributions of this work are:

- We create an open-access annotated dataset with 1,014 data points, generated from 23 publicly available wills. In addition to the three-tuple setting unique to the legal domain, this dataset also contains texts considerably longer than in other open-domain NLI datasets, such as SNLI (Bowman et al., 2015). The average length of our texts is 269 tokens, while the average lengths of premises and hypotheses in SNLI are 8 and 14 tokens, respectively.

- We demonstrate that validity assessment of legal will statements can be handled reasonably well with state-of-the-art NLI models when trained on our dataset. However, low scores in group accuracy, which is a stricter evaluation measure calculated with a group of positive and negative examples generated from the same statement as a unit, indicate that work remains before NLI models fully understand legal language.

- We explain how the trained models utilize our dataset via ablation tests that remove various

Electronic copy available at: https://ssrn.com/abstract=4279728

pieces of information at prediction time (laws, conditions, or both), and through post-hoc explainability analyses using LIME (Ribeiro et al., 2016). Our analyses indicate that in most cases the NLI classifiers use meaningful phrases from all three pieces of text, indicating the models do capture useful information. However, in some situations, the models use features that are not intuitive for humans, indicating that the task is incomplete.

2 Related Work

Natural language inference (NLI), also known as Recognizing Textual Entailment (RTE), determines entailment relations between a pair of sentences: a premise and a hypothesis. Their relationship is either: a) entailment, if a premise entails a hypothesis; b) contradiction, if a premise contradicts a hypothesis; or c) neutral, if a premise neither entails nor contradicts a hypothesis. NLI has been a key framework in natural language processing since Dagan et al. (2006) proposed the RTE challenge.

Numerous datasets exist for general NLI tasks. For example, the Stanford Natural Language Inference (SNLI) Corpus (Bowman et al., 2015) dataset contains about 570,000 pairs of sentences (premises and hypotheses) generated from Flickr image captions. More recently, datasets for domain-specific NLI tasks are being introduced. SciNLI (Sadat and Caragea, 2022) is drawn from scientific texts. It contains 107,412 sentence pairs extracted from papers in natural language processing and computational linguistics.

There have been a few NLI-related resources for the legal domain, including IFlyLegal (Wang et al., 2019), StAtutory Reasoning Assessment (SARA) dataset (Holzenberger et al., 2020), Graph-based Causal Inference (GCI) framework (Liu et al., 2021), and ContractNLI (Koreeda and Manning, 2021). One IFlyLegal module was developed for natural language article inference; it provides relevant legal articles when asked legal questions. SARA contains rules extracted from the US Internal Revenue Code and natural language questions. GCI generates causal graphs based on fact descriptions. Though the framework was not specifically designed for law, Liu et al. (2021) demonstrated that it can be utilized for legal text analysis. ContractNLI is a dataset consisting of 607 legal contracts designed to handle document-level natural language inference.

Our work differs from these existing NLI works in two ways: (a) it contains three types of input information; and (b) it operates in a pragmatic middle space with respect to text length: larger than datasets with sentence-level texts, which are insufficient to capture legal details, but shorter than document-level inference datasets.

3 Dataset

Our dataset includes three types of entries (will statements, applicable law, and facts representing external state) rather than the usual two types. Validity often depends upon such external circumstances. Thus, a will statement can be legally valid under some facts, but not others. Hence, a will can change from valid to invalid or vice versa; wills must, at minimum, be evaluated when executed (drafted and signed) and when probated (where a court determines whether to implement its provisions). For example, if a will contains a statement appointing a specific person as executor, that person's eligibility must be verified. This requires information about both the law and the person's circumstances at probate time. For example, according to Tennessee Code section 40-20-115, any person who has been imprisoned cannot be an executor. Thus, to assess eligibility of a Tennessee executor in Tennessee, one must know whether they have been in prison at probate time.

3.1 Data Collection

Our data was collected from the U.S. Wills and Probates datasets in Ancestry, which contains documents in the public domain.[1] We chose 23 wills based on three criteria: 1) whether the wills were handwritten; 2) the wills' execution and probation dates; and 3) where the wills were executed and probated. We excluded handwritten wills due to the difficulty of OCR text recognition. Execution and probation dates can affect interpretation of wills; we excluded wills executed before 1970 and probated before 2000. Lastly, we only collected wills from Tennessee, because including wills from multiple states would require analyzing each state's laws governing wills. Tennessee had the greatest number of wills meeting our criteria.

All collected wills were anonymized. Personal information was replaced with special tokens denoting the type of information (such as [Person-

[1]Court documents are in the public domain in the US.

Electronic copy available at: https://ssrn.com/abstract=4279728

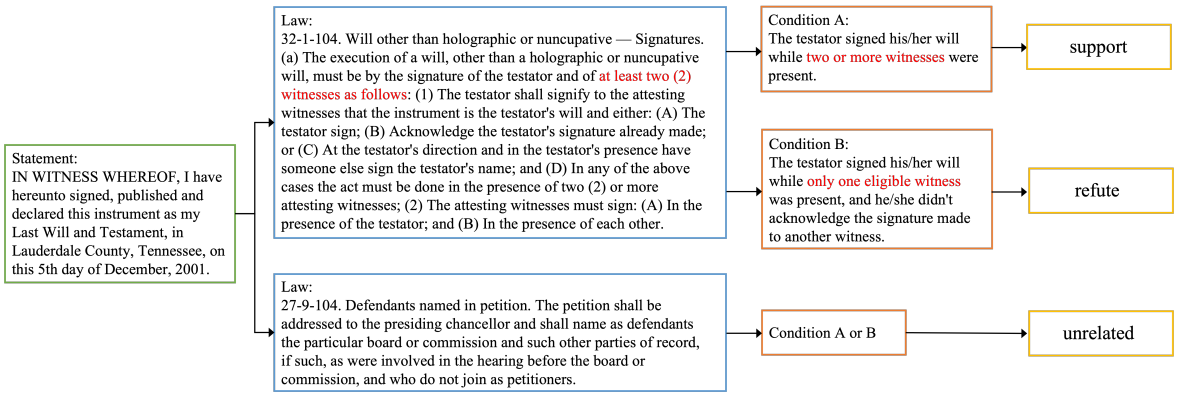


Figure 1: A demonstration of the annotation procedure. Given a law and condition, each statement is evaluated for validity. *Support*: it is supported by the law and condition provided; *Refute*: refuted by the law and condition. *Unrelated*: not relevant to the law. When a statement and a law are unrelated to each other, a condition is not required for classification. However, it is still assigned one of the conditions from either support or refute case. The goal is to prevent NLI models from depending on contextual difference (i.e., "texts without conditions are all unrelated cases") rather than features when making predictions.

n], [Address-n], [Number-n], where n identified each person or object in the will. (Suntwal et al. (2019) suggested this anonymization method). Anonymization was performed manually to prevent including personal information in the dataset.

3.2 Annotation

Each will statement was annotated as *support*, *refute*, or *unrelated* based on a given condition and law. (Our annotators set hypothetical conditions when necessary to categorize statements.)

Importantly, each statement was annotated multiple times. A statement was annotated once as support, once as refute, and thrice as unrelated.[2] This ensured that every statement was labeled with all three classes in the same ratio (support:refute:unrelated=1:1:3). Also, annotating statements three times more with "unrelated" than support and refute enabled including a greater range of laws in the dataset. Formally, our annotation procedure included the following steps:

(1) preprocessing: extracting texts from collected wills using OCR[3], and copying statements into the dataset;

(2) for each statement, identifying the Tennessee legal provision that supported, refuted, or was unrelated to the statement's validity;

(3) adding a condition specifying external circumstances relevant to validity;

(4) repeating these steps to generate five annotations per statement;

(5) anonymizing all statements by replacing personal information with special tokens.

Figure 1 demonstrates the annotation procedure with examples.

Two annotators participated in the task. One is a law student, and annotation was supervised by author Bambauer, a law professor. The other annotator does not have legal training, but ongoing discussions and reviews ensured the uniform quality. The annotators contributed equally. Annotators worked individually but shared annotation guidelines. After the annotation was complete, cross-annotation was conducted on 200 randomly selected items (100 items from each annotator's set; each annotator worked blinded on items drawn from the other annotator's set) to check inter-annotator agreement. We calculated Kappa agreement score based on the cross-annotation result, and the score is 0.91 (rounded to hundredth).

After the annotation was complete, the dataset was split into training, development, and test sets. There are 1014 data points in our dataset, split 50:25:25% (train: 504, development: 255, test: 255). When splitting, a group containing annotations from a single statement (one support, one refute, and three unrelated) was treated as a single

[2] A few exceptions occurred when:
a) it was impossible to create a case where the statement was refuted. In such cases, statements were not annotated as "refute."
b) a statement had more than one relevant legal rule. Such statements were annotated twice as support, twice as refute, and once as unrelated.

[3] The collected wills were scanned documents, so OCR was necessary to extract machine-readable text.

Electronic copy available at: https://ssrn.com/abstract=4279728

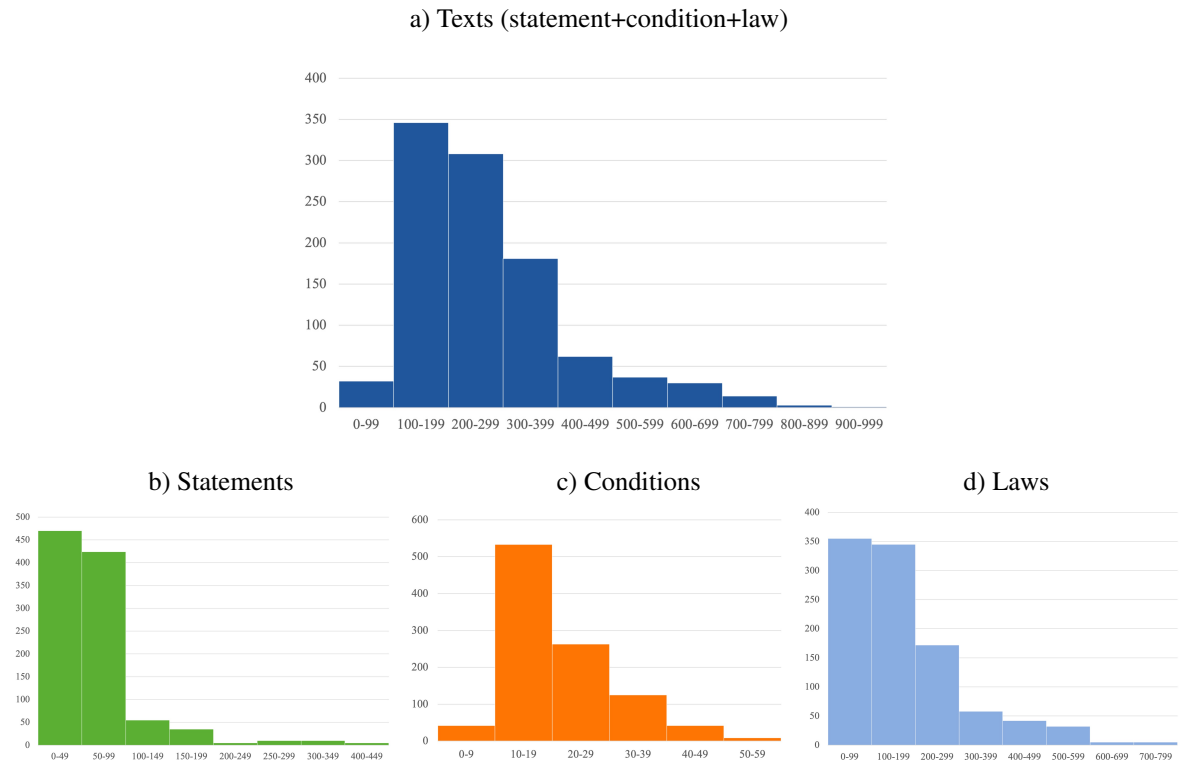


Figure 2: Histograms demonstrating the length distribution of a) full texts (i.e., statement+condition+law), b) statements, c) conditions, and d) laws in terms of token counts. Token counts are plotted on the X-axis; the Y-axis indicates the number of texts/statements/conditions/laws for each bin.

Datasets	**Texts**	**Tokens**
Train	504	135218
Development	255	69643
Test	255	68042
Total	**1014**	**272903**

Table 1: The number of texts and tokens contained in each dataset and in total

unit. Each annotator's work was equally represented in all three datasets. All 23 wills appear in more than one of the train, dev, and test partitions. However, given the nature of the will statements from the wills and how they were annotated, data leakage is unlikely. Will statements are independent and generally do not convey information about other statements from the same wills. Summary statistics are in Table 1 and Figure 2.

3.3 Characteristics of the Dataset

Our dataset has two characteristics distinguishing it from other datasets: a) texts are composed of three input types, and b) texts contain a large number of tokens.

First, our task requires a third input type: a condition, since validity often depends upon conditions.

Second, our texts tend to contain a large number of tokens. 44% of our texts contain 200 or more tokens. The average token number was 269.14; 79 texts contain more than 512 tokens (the threshold for most NLI models). Popular NLI datasets are shorter.

3.4 Implementation

We trained multiple NLI models with our dataset to assess their performance: four transformer models and four sentence-transformer models. The transformer models include bert-base-uncased (Devlin et al., 2018), distilbert-base-uncased (Sanh et al., 2019a), roberta-large-mnli (Liu et al., 2019), and longformer-base-4096 (Beltagy et al., 2020). Bert-base-uncased and dilstilbert-base-uncased were trained to set baselines on the task. In addition to the baseline models, we used roberta-large-mnli and longformer-base-4096. Among sentence-transformer models (Reimers and Gurevych, 2019), four pretrained models with top average performances (based on the Model Overview page[4]) were chosen: all-mpnet-base-v2[5], multi-qa-mpnet-base-

[4] https://www.sbert.net

[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Electronic copy available at: https://ssrn.com/abstract=4279728

dot-v1[6], all-distilroberta-v1 (Sanh et al., 2019b)[7], and all-MiniLM-L12-v2 (Wang et al., 2020)[8].

To distinguish between statements, laws, and conditions in the concatenated texts, we prefixed each with a special token: `[STATE]`, `[LAW]`, and `[COND]`.

Both transformer models and sentence-transformer models were trained on PyTorch 1.11.0 with Cuda 11.3. All transformer models were trained using the Trainer class provided by HuggingFace. Learning rates and training epochs were tuned on the development partition. Sentence-transformer models were trained with Sentence Transformer Fine-Tuning (SetFit) proposed by Wasserblat (2021). It was slightly adapted to fit the multi-class classification task, but its fundamentals remain intact. SetFit utilizes sentence pairs generated within the same class as training data. It fits the model with the data to minimize the Softmax Loss. Once the model is fitted, it is used to encode the training and development (or test, when it is a testing phase) datasets. The encoded data is used to fit the Logistic Regression classifier (with 'liblinear' solver) which, finally, makes predictions.

All models except for longformer-base-4096 were trained and evaluated on the truncated datasets due to the models' token number limitations. The models can only process texts with 512 or fewer tokens without truncation (all-MiniLM-L12-v2: 256 tokens; all-mpnet-base-v2: 384 tokens; other models, except longformer-base-4096: 512 tokens). When truncating datasets, the ratio of average token numbers among three input types (statements, laws, and conditions) was considered to prevent losing excessive information from a single type of input.

4 Results and Analysis

4.1 Evaluation Measures

We report results using standard accuracy, precision, recall, and F1 scores. To ensure all labels are well represented, precision, recall, and F1 scores were computed with Macro average. Additionally, as suggested by Elazar et al. (2021), we introduce a new measure called group accuracy (GA). GA is calculated with a group of positive and negative examples from the same statement (rather than each text) as a unit. If a group has one or more incorrect predictions, the group is incorrect. If a model correctly understands a will statement, it should perform equally well on all examples from the same statement.

4.2 Results from Trained Models

Table 2 shows the models' performances on our test partition. Overall, the models demonstrated good performances. Each achieved more than 80% in all metrics but group accuracy. Roberta-large-mnli showed the best performance. It achieved over 96% in all metrics but group accuracy (84.31%), suggesting it can handle the task reasonably well. However, the large difference between accuracy and group accuracy suggests that the models' understanding of the task is superficial.

Accuracy for the *unrelated* label was higher than for *support* and *refute* in all models except all-mpnet-base-v2 and multi-qa-mpnet-base-dot-v1. This higher accuracy for *unrelated* label may be partially attributable to the gap between accuracy and group accuracy, as it would inflate overall accuracy. However, since the gap between accuracy and group accuracy is not significantly smaller for all-mpnet-base-v2 and multi-qa-mpnet-base-dot-v1 (i.e., the models where accuracy for *unrelated* was not higher than the other labels), it is more likely that the gap originated from the strictness of group accuracy and the models' superficial understanding of will statements.

Further, roberta-large-mnli (trained with truncated inputs) showed better performance than longformer-base-4096 (trained with full length inputs). This suggests that models with token number limitations can still perform well with long inputs when truncated properly.

4.3 Experiment with Different Input Lengths

To investigate whether the models' performance was affected by input length and/or input truncation, we varied input lengths. Inputs were classified into three categories by length: a) short (equal or less than 192 or 256 tokens); b) regular (larger than 192 or 256 tokens and equal or less than 384 or 512); and c) long (larger than 384 or 512)[9]. The

[6] https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1

[7] https://huggingface.co/sentence-transformers/all-distilroberta-v1

[8] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

[9] For all-mpnet-base-v2 model, 192 and 384 were used as dividing points instead of 256 and 512, as its maximum token length is 384.

Electronic copy available at: https://ssrn.com/abstract=4279728

Model	Precision	Recall	F1	Accuracy	GA
Transformers models					
bert-base-uncased	81.47	82.20	81.81	87.06	49.02
distilbert-base-uncased	82.37	82.59	82.48	87.06	50.98
longformer-base-4096	94.09	93.69	93.85	94.90	74.51
roberta-large-mnli	**96.67**	**96.25**	**96.42**	**96.86**	**84.31**
Sentence-Transformers models					
all-MiniLM-L12-v2	81.50	84.54	82.64	85.49	50.98
all-disilroberta-v1	83.43	85.99	84.22	87.06	49.02
multi-qa-mpnet-base-dot-v1	91.48	94.13	92.65	93.73	74.51
all-mpnet-base-v2	92.93	95.05	93.82	94.90	76.47

Table 2: Results of all classifiers trained on our dataset using five measures: precision, recall, F1, accuracy, and group accuracy (GA in table, see Section 4.1). Precision, recall, and F1 scores were computed with Macro average. Group accuracy is calculated with positive and negative examples generated from the same statement as a unit. If a group contains one or more incorrect predictions, the group is considered incorrect.

Model	Precision	Recall	F1	Accuracy
all-mpnet-base-v2 (sample n = 52)				
token <= 192	93.75	93.15	92.91	94.23
192 < token <= 384	90.48	89.10	89.63	94.23
384 < token	86.35	89.49	87.08	88.46
longformer-base-4096 (sample n = 20)				
token <= 256	95.24	88.89	90.77	95.00
256 < token <= 512	80.56	71.67	70.93	80.00
512 < token	73.81	66.67	66.93	75.00
roberta-large-mnli (sample n = 20)				
token <= 256	100	100	100	100
256 < token <= 512	94.87	85.00	88.76	90.00
512 < token	94.44	93.33	93.27	95.00

Table 3: Results with different input lengths. Models generally performed worse when token numbers in inputs were larger. Roberta-large-mnli performed better than longformer-base-4096 in predicting long inputs (token > 512) despite truncation. Roberta-large-mnli also showed better performance on long inputs than regular inputs, indicating input truncation did not worsen the model's performance.

number of inputs in each category varied. To control the impact of varying sample size, categories with larger sample sizes were reduced by random sampling to match the smallest sample size. For this experiment, we used the three models with best performance on the full dataset.

Table 3 shows the results. Models generally performed worse with larger input token numbers. All three models performed best with inputs with smaller token numbers. One interesting finding is that roberta-large-mnli performed better than longformer-base-4096 in predicting inputs with large token numbers (n > 512) despite truncation. Roberta-large-mnli showed better performance on long inputs than regular inputs, indicating truncation did not negatively affect the model's performance. This finding aligns with the observation from the overall results that models with token number limitation can still show good performances on long inputs when properly truncated. However, results differed with a smaller token number limitation. All-mpnet-base-v2 performed worst on long inputs (where truncation occurred), indicating truncation negatively impacted performance.

4.4 Results from Ablation Experiment

To determine whether the models correctly rely on all three types of texts (statements, conditions, and laws), we conducted an ablation experiment, training the models with datasets lacking one (law or condition) or two types (law+condition) of inputs. Poliak et al. (2018) suggested a similar experiment

Electronic copy available at: https://ssrn.com/abstract=4279728

Model	Precision	Recall	F1	Accuracy	GA
Statements and laws					
all-mpnet-base-v2	75.64	76.47	75.98	81.96	27.45
longformer-base-4096	75.72	75.98	75.79	82.75	37.25
roberta-large-mnli	77.93	77.51	77.55	84.31	39.22
Statements and conditions					
all-mpnet-base-v2	46.67	57.13	44.80	44.31	0.00
longformer-base-4096	58.00	49.44	51.30	58.82	0.00
roberta-large-mnli	37.63	41.72	38.08	56.86	0.00
Statements only					
all-mpnet-base-v2	36.76	35.94	34.04	36.47	0.00
longformer-base-4096	18.17	33.33	23.52	54.51	0.00
roberta-large-mnli	18.17	33.33	23.52	54.51	0.00

Table 4: Results from the ablation experiment where three models were trained with datasets lacking one (law or condition) or two types (law+condition) of inputs. Results demonstrate that the models' performances deteriorate if one or more input type(s) is omitted. Thus, both types of inputs affect the models' performances, though laws have larger impact than conditions.

to this one to set hypothesis-only baselines. We used the three models with the best performances on the full dataset for this experiment.

Table 4 shows the results. Performance deteriorated if any input type was omitted. Among the models trained on partial data, those trained with statements and laws performed better, with F1 scores over 75%. However, group accuracy dropped to the 30s or even 20%s, indicating actual understanding decreased considerably. Models trained with statements and conditions performed significantly worse. Results ranged between 38% and 52% for F1, and group accuracy dropped to 0. Models trained only with statements performed worst, with F1 scores dropping to 35% or lower.

This degradation from partial data shows including both conditions and laws positively affects the models' performance. Models use all three types of inputs when making predictions[10]. Laws have more impact on performance than conditions, since deterioration in results without laws was more significant than from models trained without conditions. This is realistic, since laws set the parameters within which conditions operate to make a given provision valid or invalid.

[10] A reviewer suggested an experiment where laws and conditions are packed into premises and treat the task as normal NLI. A pilot experiment with the best performing model (roberta-large-mnli) found a slight deterioration in performance when laws and conditions were combined as premises (precision: 93.09, recall: 93.89, F1: 93.41, accuracy: 94.12, GA: 74.51; see Table 2 for comparison with main result). The biggest difference came from GA score (74.51 vs. 84.31). This lowered GA score implies that the model's understanding worsened when conditions and laws were combined into premises. Given this preliminary result, it is likely that three-way distinction in input types positively contributes to the model's performance. Nevertheless, as the performance difference is not significantly large, further investigation is needed.

4.5 Understanding Results with LIME

To clarify model behavior, we used Local Interpretable Model-Agnostic Explanations (LIME) to explain our classifiers' predictions (Ribeiro et al., 2016). We implemented LIME on predictions made by the best performing model (roberta-large-mnli, fine-tuned on our data). The LIME results revealed that the model tends to correctly rely on all three texts (statements, laws, conditions) for a prediction, and uses meaningful features in many cases. However, sometimes the model utilized features that are not intuitive for humans.

4.5.1 A Correct Example

Figure 3 shows a case where the model made a correct prediction based on features sensible to humans. Figure 3's text is a statement saying two or more witnesses witnessed the testator signing the will and signed in each other's presence. It includes a condition stating that one of the two witnesses was ineligible to serve under Tennessee law, and the law specifying witness eligibility in Tennessee. The given condition and the law invalidate the will statement, rendering it as *refute*. The model correctly predicted the outcome based on tokens such as *ineligible*, *witnesses*, *One*, *testament*, and *2*. The top two tokens with greatest impact on the model's prediction were *ineligible* and *witnesses*, with importance score of 0.60 and 0.14, respectively. It is

Electronic copy available at: https://ssrn.com/abstract=4279728

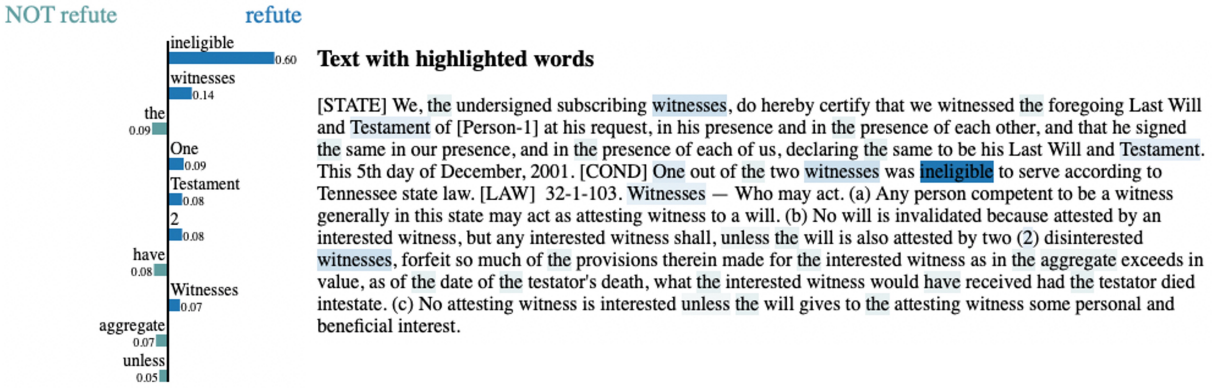


Figure 3: Examples of LIME explanations showing a case where the model (roberta-large-mnli) makes correct predictions based on features sensible to humans.

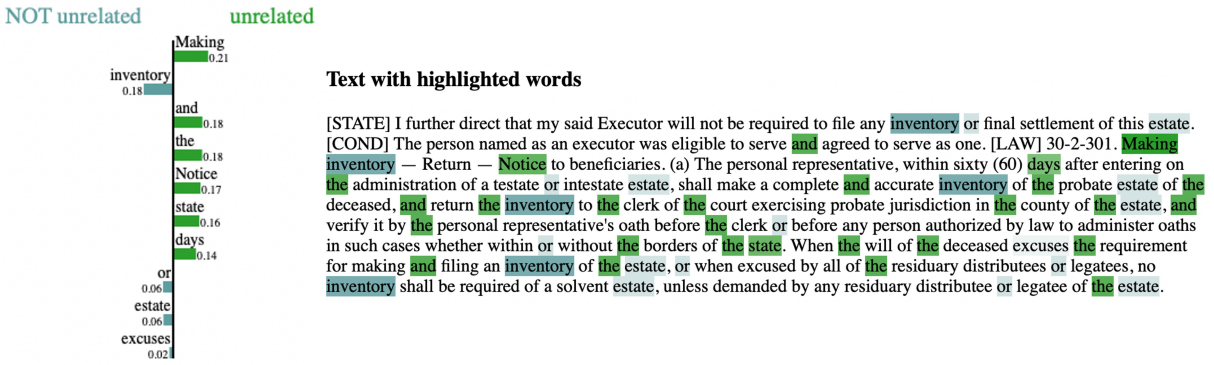


Figure 4: An example of a LIME explanation showing a case where the model (roberta-large-mnli) makes an incorrect prediction based on the tokens such as *Making*, *and*, and *the* which bear little semantic relevancy to the gist of the text.

understandable that these tokens have high scores, as they are the keywords ("**One** out of two (**2**) **witnesses** is **ineligible**") which provide grounds for revoking a will statement.

4.5.2 An Incorrect Example

Figure 4 shows a case where the model made an incorrect prediction based on irrelevant tokens. The intent of the will statement is to excuse the Executor from filing inventory. This is not contradicted by the condition and law, requiring the model to classify the text as *support*. Nevertheless, the model incorrectly classified it as *unrelated*. The reason for the incorrect prediction is that the model (probably) relied on irrelevant tokens such as *Making*, *and*, and *the*, which bear little semantic relevancy to the text, versus more relevant tokens such as *inventories* and *excuses* when making the prediction. This is likely due to overfitting induced by longer texts. This LIME explanation shows that even the best performing model still has room for improvement.

5 Future Work

This work can be expanded in several directions. First, it can be extended to cover multiple states by adapting the models or adding more data. Future work can investigate if models trained on a single state's data can be adapted to evaluate data from other states. Also, wills from multiple states could be added to the dataset. We expect the augmented dataset would enhance the models' performance on evaluating wills from other states.

This work can be expanded to other legal domains. Models trained on our data can be adapted to similar tasks in other legal domains.

Lastly, this work can be extended by investigating novel transformer models. Given the uniqueness of our dataset with regard to the number of inputs and text length, it is likely that further experiment and modification is needed to handle such characteristics.

Electronic copy available at: https://ssrn.com/abstract=4279728

6 Conclusion

This work presented an annotated dataset for natural language inference (NLI) in the legal domain, consisting of 1,014 data points generated from 23 publicly available wills. The dataset is novel for two reasons: it included texts with three input types (statement, law, and condition) rather than two (premise and hypothesis) in the traditional NLI, and included texts are longer than in general NLI datasets. The NLI models trained on our dataset showed reasonable performance in assessing the validity of will statements. Ablative experiments demonstrated that the models' performances worsen if any input type (condition, law, or both) is omitted. This suggests that the models utilize all three input types. The LIME implementation reveals that even the best performing model makes errors in some cases by using semantically irrelevant tokens. Our open-access dataset is publicly available at: https://github.com/ml4ai/nli4wills-corpus

Limitations

Our dataset consists of a relatively small number of data points (1,014 texts). Annotating will statements with relevant laws and conditions is a highly demanding, time intensive task. A larger size of dataset would likely further improve the models' performances.

Our dataset only includes wills executed and probated in Tennessee, with execution after 1970 and probate after 2000. Due to these restrictions, our framework might produce incorrect results on inputs from wills from different settings. Supplementing the dataset with wills from more diverse settings would address this limitation. Even though the scope is limited to a single state (Tennessee), our study demonstrates that transformer models trained on the dataset can evaluate the validity of statements from wills with reasonable accuracy. Future additions to our dataset will be available at the same URL: https://github.com/ml4ai/nli4wills-corpus

This work does not involve humans in the loop. Considering how crucial accuracy is for the task (i.e., legal validity evaluation), the work would have benefited much from involving domain experts. Even though our study discovered that the state-of-the-art transformer models can show good performances (over 85% accuracy in all 8 models) without human interaction, it was also found that the models' understanding on the task is rather superficial (GAs ranging from 49-85% in Table 2). Including humans-in-the-loop could be a solution for enhancing the models' understanding on the task.

Ethics Statement

We collected legal wills from Ancestry as a part of our dataset creation process. The wills probated in court are in the public domain in the US, and we did not violate Ancestry's Terms and Conditions when collecting wills. We also anonymized the wills by replacing any personal identifiable information contained in the documents with special tokens.

Our datasets and codes are released to public. We believe our released datasets and codes would contribute to society by promoting further NLI endeavors in the legal domain. The resources could potentially assist with people reviewing wills, but they should not be considered as legal advice. To avoid any confusion, we placed a disclaimer that the users must not rely on any information provided from our resources when making legal decisions and should instead consult with an attorney.

Acknowledgements

We thank the reviewers for their thoughtful comments and suggestions. This work was partially supported by the National Science Foundation (NSF) under grant #2217215, and by University of Arizona's Provost Investment Fund. Mihai Surdeanu and Clayton Morrison declare a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification,*

Electronic copy available at: https://ssrn.com/abstract=4279728

and Recognising Tectual Entailment, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

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.

Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. 2021. Back to square one: Artifact detection, training and commonsense disentanglement in the Winograd schema. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10486–10500, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering. *CoRR*, abs/2005.05257.

Yuta Koreeda and Christopher D. Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. *CoRR*, abs/2110.01799.

Xiao Liu, Da Yin, Yansong Feng, Yuting Wu, and Dongyan Zhao. 2021. Everything has a cause: Leveraging causal inference in legal text analysis. *CoRR*, abs/2104.09420.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Mobashir Sadat and Cornelia Caragea. 2022. SciNLI: A corpus for natural language inference on scientific text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7399–7409, Dublin, Ireland. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019a. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019b. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Sandeep Suntwal, Mithun Paul, Rebecca Sharp, and Mihai Surdeanu. 2019. On the importance of delexicalization for fact verification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3413–3418, Hong Kong, China. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Ziyue Wang, Baoxin Wang, Xingyi Duan, Dayong Wu, Shijin Wang, Guoping Hu, and Ting Liu. 2019. IFlyLegal: A Chinese legal system for consultation, law searching, and document analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 97–102, Hong Kong, China. Association for Computational Linguistics.

Moshe Wasserblat. 2021. Sentence transformer fine-tuning (setfit): Outperforming gpt-3 on few-shot text-classification while being 1600 times smaller.

Electronic copy available at: https://ssrn.com/abstract=4279728

APPENDIX 1011H

web.archive.org/web/20221128060841/https://papers.ssrn.com/sol3/papers.cfm?abstract_id=42
Sign in
INTERNET ARCHIVE Wayback Machine
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4279728
Go
8 captures
28 Nov 2022 - 30 May 20
OCT NOV JAN
28
2021 2022 2024
About this capture

Download This Paper | Open PDF in Browser
Add Paper to My Library
Share:
Feedback

Validity Assessment of Legal Will Statements as Natural Language Inference

Findings of the Association for Computational Linguistics: EMNLP 2022 (forthcoming 2022)
Arizona Legal Studies Discussion Paper No. 22-21

11 Pages • Posted: 17 Nov 2022 • Last revised: 21 Nov 2022

Alice Kwak
The University of Arizona - Department of Linguistics

Jacob Israelsen
University of Arizona - James E. Rogers College of Law

Clayton Morrison
University of Arizona - School of Information

Derek E. Bambauer
University of Arizona - James E. Rogers College of Law

Do you have a job opening that you would like to promote on SSRN?
Place Job Opening

Paper statistics
DOWNLOADS 22
ABSTRACT VIEWS 81
Loading citations...
PlumX Metrics

The Wayback Machine - https://web.archive.org/web/20221128060841/https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4279728

Download This Paper (Delivery.cfm/SSRN_ID4279728_code575020.pdf?abstractid=4279728&mirid=1) Open PDF in Browser (Delivery.cfm/SSRN_ID4279728_code575020.pdf?abstractid=4279728&mirid=1&type=2)

Add Paper to My Library

Share:

Validity Assessment of Legal Will Statements as Natural Language Inference

Findings of the Association for Computational Linguistics: EMNLP 2022 (forthcoming 2022)
Arizona Legal Studies Discussion Paper No. 22-21

11 Pages
Posted: 17 Nov 2022
Last revised: 21 Nov 2022

Alice Kwak (https://web.archive.org/web/20221128060841/https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=5582596)
The University of Arizona - Department of Linguistics

Jacob Israelsen (https://web.archive.org/web/20221128060841/https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=5582711)
University of Arizona - James E. Rogers College of Law

Clayton Morrison (https://web.archive.org/web/20221128060841/https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=5582732)
University of Arizona - School of Information

Derek E. Bambauer (https://web.archive.org/web/20221128060841/https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=441439)
University of Arizona - James E. Rogers College of Law

Mihai Surdeanu (https://web.archive.org/web/20221128060841/https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=5582734)
University of Arizona - Department of Computer Science

Date Written: November 15, 2022

Abstract

This work introduces a natural language inference (NLI) dataset that focuses on the validity of statements in legal wills. This dataset is unique because: (a) each entailment decision requires three inputs: the statement from the will, the law, and the conditions that hold at the time of the testator's death; and (b) the included texts are longer than the ones in current NLI datasets. We trained eight neural NLI models in this dataset. All the models achieve more than 80% macro F1 and accuracy, which indicates that neural approaches can handle this task reasonably well. However, group accuracy, a stricter evaluation measure that is calculated with a group of positive and negative examples generated from the same statement as a unit, is in mid 80s at best, which suggests that the models' understanding of the task remains superficial. Further ablative analyses and explanation experiments indicate that all three text segments are used for prediction, but some decisions rely on semantically irrelevant tokens. This indicates that overfitting on these longer texts likely happens, and that additional research is required for this task to be solved.

Keywords: natural language, natural language inference, NLI, machine learning, smart contracts, blockchain, AI, artificial intelligence, algorithm, automation, texts, legal wills

Suggested Citation >

Show Contact Information >

Download This Paper (Delivery.cfm/SSRN_ID4279728_code575020.pdf?abstractid=4279728&mirid=1) Open PDF in Browser (Delivery.cfm/SSRN_ID4279728_code575020.pdf?abstractid=4279728&mirid=1&type=2)

Do you have a job opening that you would like to promote on SSRN?

Place Job Opening (https://web.archive.org/web/20221128060841/https://www.ssrn.com/index.cfm/en/Announcements-Jobs/)

Paper statistics

DOWNLOADS 22

ABSTRACT VIEWS 81

Loading citations...

PlumX Metrics
(https://web.archive.org/web/20221128060841/https://plu.mx/ssrn/a/?ssrn_id=4279728)

Related eJournals

University of Arizona James E. Rogers College of Law Legal Studies Research Paper Series (https://web.archive.org/web/20221128060841/https://papers.ssrn.com/sol3/JELJOUR_Results.cfm?form_name=journalBrowse&journal_id=872768)

Follow

Wills, Trusts, & Estates Law eJournal (https://web.archive.org/web/20221128060841/https://papers.ssrn.com/sol3/JELJOUR_Results.cfm?form_name=journalBrowse&journal_id=677861)

Follow

View more >

Feedback

Submit a Paper > (https://web.archive.org/web/20221128060841/https://hq.ssrn.com/submissions/CreateNewAbstract.cfm)

SSRN Quick Links

SSRN Rankings

About SSRN

(https://web.archive.org/web/20221128060841/https://www.facebook.com/SSRNcommunity/)
(https://web.archive.org/web/20221128060841/https://www.linkedin.com/company/493409?trk=tyah&trkInfo=clickedVertical%3Acompany%2CentityType%3AentityHistoryName%2CclickedEntityId%3Acompany_493409%2Cidx%3A0)
(https://web.archive.org/web/20221128060841/https://twitter.com/SSRN)

(https://web.archive.org/web/20221128060841/https://www.elsevier.com/)

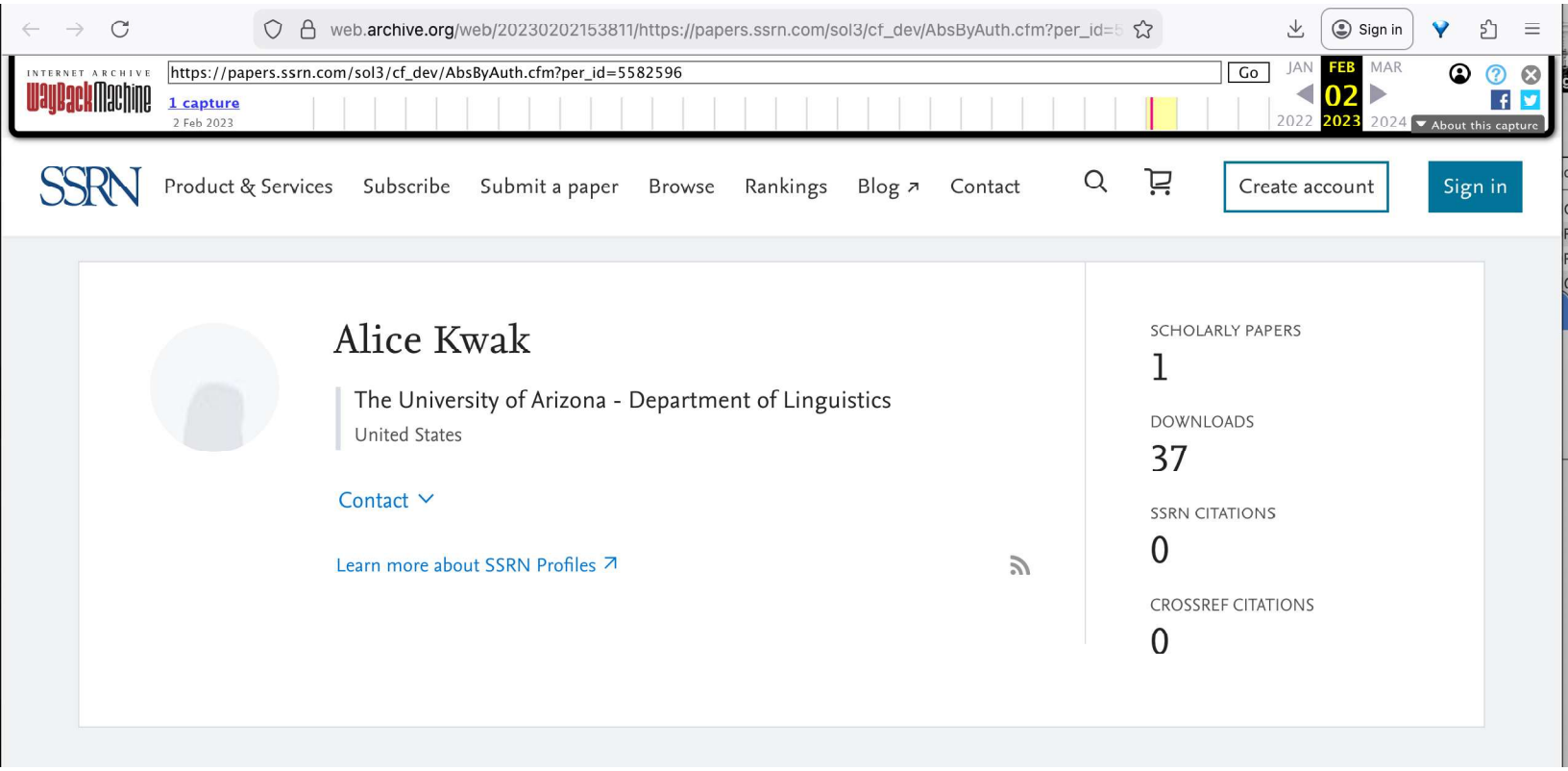
Copyright (https://web.archive.org/web/20221128060841/https://www.ssrn.com/index.cfm/en/dmca-notice-policy/) Terms and Conditions (https://web.archive.org/web/20221128060841/https://www.ssrn.com/index.cfm/en/terms-of-use/)
Privacy Policy (https://web.archive.org/web/20221128060841/https://www.elsevier.com/legal/privacy-policy)

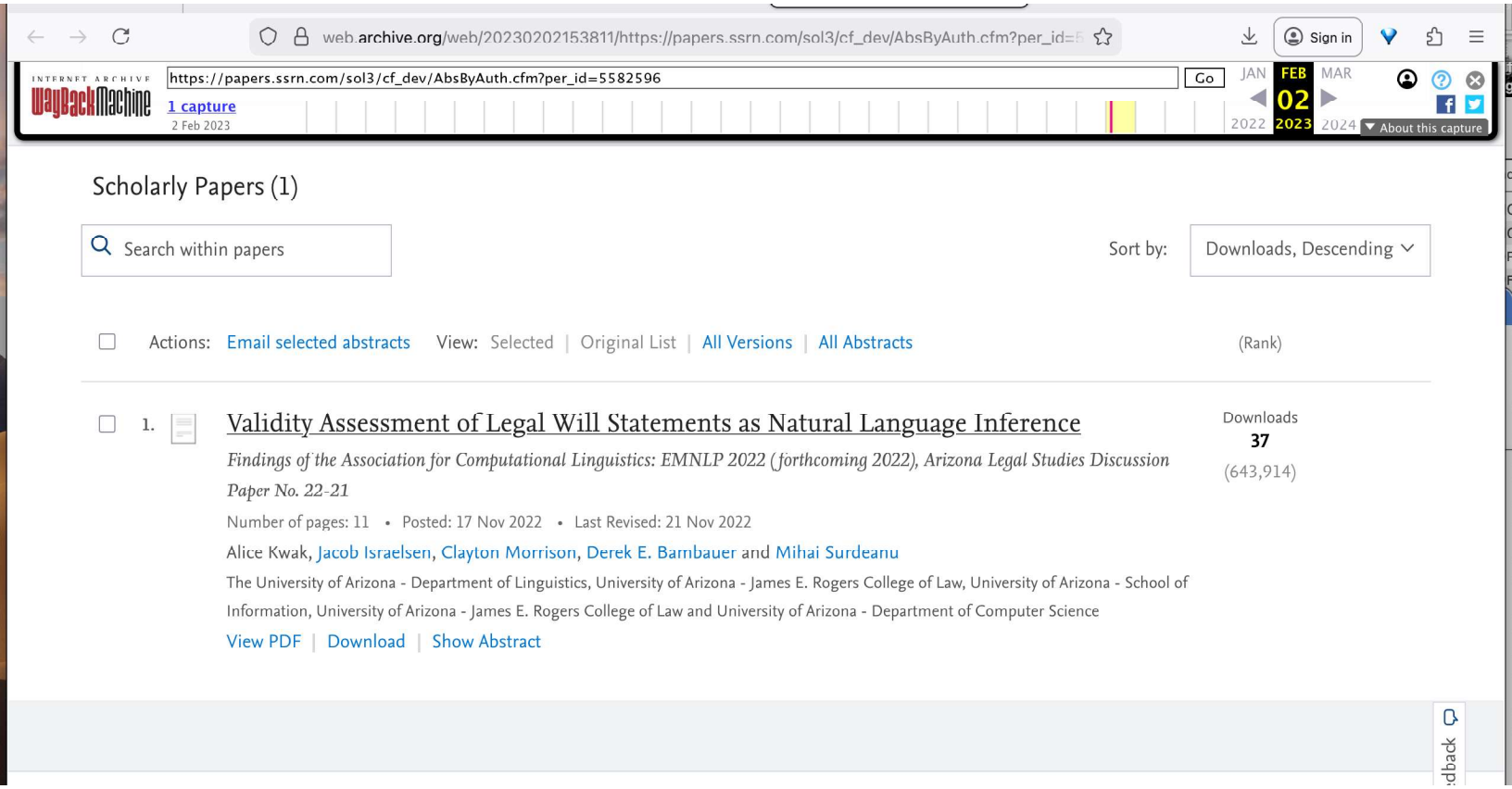
We use cookies to help provide and enhance our service and tailor content.

To learn more, visit Cookie Settings.

(https://
web.archive.org/
web/20221128060841/
http://
www.relx.com/)

(https://web.archive.org/web/20221128060841/https://papers.ssrn.com/sol3/updateInformationLog.cfm?process=true)


web.archive.org/web/20230202153811/https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=5
Sign in
INTERNET ARCHIVE
WaybackMachine
https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=5582596
Go
1 capture
2 Feb 2023
JAN
FEB
MAR
02
2022
2023
2024
About this capture
SSRN
Product & Services
Subscribe
Submit a paper
Browse
Rankings
Blog
Contact
Create account
Sign in
Alice Kwak
The University of Arizona - Department of Linguistics
United States
Contact
Learn more about SSRN Profiles
SCHOLARLY PAPERS
1
DOWNLOADS
37
SSRN CITATIONS
0
CROSSREF CITATIONS
0


web.archive.org/web/20230202153811/https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=5
Sign in
INTERNET ARCHIVE
WaybackMachine
https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=5582596
Go
1 capture
2 Feb 2023
JAN
FEB
MAR
02
2022
2023
2024
About this capture
Scholarly Papers (1)
Search within papers
Sort by:
Downloads, Descending
Actions:
Email selected abstracts
View:
Selected
Original List
All Versions
All Abstracts
(Rank)
1.
Validity Assessment of Legal Will Statements as Natural Language Inference
Findings of the Association for Computational Linguistics: EMNLP 2022 (forthcoming 2022), Arizona Legal Studies Discussion Paper No. 22-21
Number of pages: 11 • Posted: 17 Nov 2022 • Last Revised: 21 Nov 2022
Alice Kwak, Jacob Israelsen, Clayton Morrison, Derek E. Bambauer and Mihai Surdeanu
The University of Arizona - Department of Linguistics, University of Arizona - James E. Rogers College of Law, University of Arizona - School of Information, University of Arizona - James E. Rogers College of Law and University of Arizona - Department of Computer Science
View PDF
Download
Show Abstract
Downloads
37
(643,914)

web.archive.org/web/20230202153811/https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=5
Sign in
INTERNET ARCHIVE Wayback Machine
https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=5582596
Go
1 capture
2 Feb 2023
JAN FEB MAR
2022 02 2023 2024
About this capture
Feedback

Submit a Paper >

Section 508 Text Only Pages

SSRN Quick Links

- SSRN Solutions
- Research Paper Series
- Conference Papers
- Partners in Publishing
- Jobs & Announcements
- Special Topic Hubs

SSRN Rankings

- Top Papers
- Top Authors
- Top Organizations

About SSRN

- SSRN Objectives
- Network Directors
- Presidential Letter
- Announcements
- Contact us
- FAQs

ELSEVIER

Copyright Terms and Conditions Privacy Policy

We use cookies to help provide and enhance our service and tailor content.

To learn more, visit Cookie settings | Your Privacy Choices.

RELX™

The Wayback Machine - https://web.archive.org/web/20230202153811/https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=5582596

Alice Kwak

The University of Arizona - Department of Linguistics

United States

Contact

- The University of Arizona - Department of Linguistics
  Email

Learn more about SSRN Profiles ↗

SCHOLARLY PAPERS
1

DOWNLOADS
37

SSRN CITATIONS
0

CROSSREF CITATIONS
0

Scholarly Papers (1)

Search within papers

Sort by:

Downloads, Descending

Actions:

Email selected abstracts View: Selected | Original List | All Versions | All Abstracts

(Rank)

1.

Validity Assessment of Legal Will Statements as Natural Language Inference

Findings of the Association for Computational Linguistics: EMNLP 2022 (forthcoming 2022), Arizona Legal Studies Discussion Paper No. 22-21

Number of pages: 11

Posted: 17 Nov 2022

Last Revised: 21 Nov 2022

Alice Kwak, Jacob Israelsen, Clayton Morrison, Derek E. Bambauer and Mihai Surdeanu

The University of Arizona - Department of Linguistics, University of Arizona - James E. Rogers College of Law, University of Arizona - School of Information, University of Arizona - James E. Rogers College of Law and University of Arizona - Department of Computer Science

Downloads: **37**
(643,914)

View PDF | Download | Show Abstract

Abstract:

Loading...

natural language, natural language inference, NLI, machine learning, smart contracts, blockchain, AI, artificial intelligence, algorithm, automation, texts, legal wills

Feedback

Submit a Paper >

SSRN Quick Links

SSRN Rankings

About SSRN

Copyright Terms and Conditions Privacy Policy

We use cookies to help provide and enhance our service and tailor content.
To learn more, visit Cookie Settings.

APPENDIX 1011I

ScienceDirect

AI Open

Volume 3, 2022, Pages 208-212

CAILIE 1.0: A dataset for Challenge of AI in Law - Information Extraction V1.0

Yu Cao, Yuanyuan Sun, Ce Xu, Chunnan Li, Jinming Du, Hongfei Lin

Department of Computer Science and Technology, Dalian University of Technology, China

Received 20 June 2022, Revised 27 November 2022, Accepted 1 December 2022, Available online 5 December 2022, Version of Record 6 December 2022.

What do these dates mean?

Show less

Outline | Share Cite

https://doi.org/10.1016/j.aiopen.2022.12.002

Get rights and content

Under a Creative Commons license

Open access

Abstract

Legal information extraction requires identifying and classifying legal elements from specific legal documents. Considering that information extraction is mainly regarded as the first step in natural language understanding, the quality of legal information extraction results certainly has an immense impact on the performance of various legal artificial intelligence (AI) downstream tasks. However, Chinese judicial information extraction datasets are very scarce due to the particularity of legal documents. In response to this situation, we constructed a dataset for Challenge of AI in Law - Information Extraction V1.0 (CAILIE 1.0). The following two features of CAILIE are worth highlighting: 1) the entity definition focuses on more fine-grained theft document information, providing more interpretability for downstream legal AI; and 2) we define entity labels with judicial attributes based on natural attribute labels to meet the needs of Chinese judicial practice. We implement some classic models on this dataset. The experimental results show that legal

information extraction is still challenging and additional research is required for this task to be solved.

Previous Next

Keywords

Legal information extraction; Legal artificial intelligence

1. Introduction

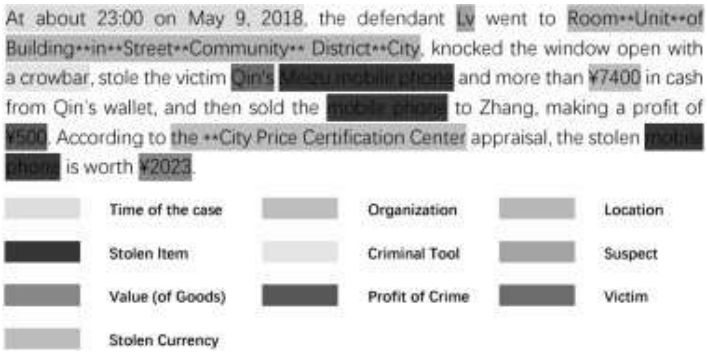
Legal information extraction is the process of extracting and classifying nouns and phrases with judicial characteristics from various legal documents, such as those pertaining to stolen items and suspects. As a subtask of legal artificial intelligence, legal information extraction is the basis for downstream applications, such as similar case retrieval, automatic sentencing advice, and legal questions and answers. Reasonably defined extracted information can be used to assist in solving downstream tasks. At the same time, the information enhances their interpretability, which is of great significance in the judicial field.

Based on general domain information extraction methods (Kuru et al., 2016, Luoma and Pyysalo, 2020, Li et al., 2021, Zhou et al., 2022), researchers adopted ontologies(Zhang et al., 2017), global consistency (Yin et al., 2018), pointer generator network(Skylaki et al., 2020) and pre-trained language model to understand long documents (Xiao et al., 2021) according to the characteristics of legal texts to improve the performance of legal information extraction. Chen et al. (2020) attempted to exploit a multi-task learning framework to extract entities and semantic relations jointly in drug-related criminal judgment.Wang et al. (2021) proposed a bi-directional feedback clause-element relation network using fine-grained contract elements and large extraction zones. However, the lack of Chinese legal information extraction datasets limits further research.

Faced with this status quo, we propose CAILIE, a Chinese legal information extraction dataset for larceny. CAILIE has two features worth highlighting: (1) the dataset contains fine-grained information for specific crimes, providing more interpretability for downstream tasks; and (2) considering the needs of judicial practice, the entity categories of this dataset are designed from the perspective of conviction. Emphasizing judicial attributes rather than natural attributes means that more legal knowledge can be obtained, which is more conducive to downstream applications such as sentencing prediction.

CAILIE contains 5247 sentences, 10 types of entities, and 27031 entities. Therefore, it is capable of providing sufficient supervision signals and reliable evaluation results. We conduct some

experiments with widely-used models on this dataset. The experimental results confirm that legal information extraction remains challenging and still needs improvement.



Download: Download high-res image (413KB)

Download: Download full-size image

Fig. 1. Sample case from CAILIE.

2. Related work

2.1. Legal information extraction

The explosive growth of public legal data has made legal information extraction an urgent problem that needs to be solved by judicial artificial intelligence. In addition to public domain information extraction methods(Nie et al., 2020, Zhou et al., 2019, Wang et al., 2022, Kuru et al., 2016, Luoma and Pyysalo, 2020, Li et al., 2021, Zhou et al., 2022), the specific writing style and wording features of legal documents have drawn significant attention. Researchers have attempted to use ontologies(Zhang et al., 2017), global consistency(Yin et al., 2018), pointer generator network(Skylaki et al., 2020) and bi-directional feedback network(Wang et al., 2021) for named entity recognition in legal documents. Notably, with the rapid development of natural language processing,Iftikhar et al. (2019) introduced machine learning algorithms to extract information from criminal judgments of the Lahore High Court. Researchers have also conducted some explorations in Chinese legal information extraction.Gong et al. (2019) studied the extraction of basic information from Chinese legal documents, such as names, criminal charges, evidence, and sentence duration. (Xiao et al., 2021) trained a pre-trained language model for understanding long Chinese legal documents.

2.2. Legal artificial intelligence

With the explosive growth of public legal data, legal artificial intelligence has attracted substantial attention. In addition to legal information extraction mentioned above, legal judgment prediction (Yang et al., 2019, Wang and Jin, 2020, Bibal et al., 2020, Niklaus et al., 2021, Huang et al., 2021, Ma et al., 2021), similar case retrieval(Jackson et al., 2003, Zeng et al., 2005, Mandal et al., 2017, Tran et

al., 2020, Westermann et al., 2021, Shao et al., 2021, Bithel and Malagi, 2021), legal question answering (Do et al., 2017, Duan et al., 2019a, Wehnert et al., 2019, Zhong et al., 2020, Kien et al., 2020, Nguyen et al., 2020) , legal document summarization(Kanapala et al., 2019, Duan et al., 2019b, Nguyen et al., 2021, Bhattacharya et al., 2021), and others are typical applications of legal artificial intelligence. As a basic task, legal information extraction can improve the performance of downstream tasks, and the extracted symbols can provide interpretability for legal applications.

2.3. Datasets in law

Given the important societal role of law, along with the high cost of legal expertise, the utilization of natural language processing to help solve legal-related problems has elicited considerable attention. Numerous datasets have been proposed for legal artificial intelligence tasks. Holzenberger et al. (2020) andKwak et al. (2022) introduced datasets to investigate the performance of natural language understanding approaches on statutory reasoning.Habernal et al., 2022, Shen et al., 2022 and Aumiller et al. (2022) aimed to fill the gap in legal document summarization by presenting an argument mining dataset, multiple target summary dataset and multi-lingual summary dataset.GilSemo et al. (2022) released a legal judgment prediction dataset focused on class action cases in the US.

To our knowledge, the techniques of named entity recognition in the Chinese legal domain still lack annotation data. The coverage rate of crimes is less than 2%, far from meeting the demand of judicial practice. To this end, we propose a Chinese legal information extraction dataset for larceny called CAILIE, which contains more than 27,000 judicial attribute entities, thereby providing rich and reliable supervision information for information extraction tasks.

In the field of named entity recognition,Dozier et al. (2010) divided entities in legal texts into judges, attorneys, companies, jurisdictions, and courts. While Cardellino et al. (2017) noted fine-grained entities unique to legal documents, they did not view them from the perspective of prosecutors. Compared with existing legal information extraction datasets, CAILIE was constructed according to Article 264 of the Criminal Law of the People's Republic of China and the relevant judicial interpretations: *Whoever steals a relatively large amount of public or private property, commits theft many times, commits a burglary or carries a lethal weapon to steal or pickpockets shall be sentenced to imprisonment of not more than 3years, limited incarceration or probation and/or a fine; if the amount involved is huge or there is any other serious circumstance, shall be sentenced to imprisonment of not less than 3years but not more than 10years and a fine; or if the amount involved is especially huge or there is any other especially serious circumstance, shall be sentenced to imprisonment of not less than 10years or life imprisonment and a fine or forfeiture of property.* In response to the factors affecting sentencing mentioned in the above legal articles, we defined 10fine-grained classes.

3. Dataset

Typically, there are six natural attribute entities, namely, Person, Location, Time, Item, Currency, and Organization. For larceny, whethe the house was entered illegally, theft with lethal weapons, pickpocketing, and others, are usually issues of concern for judicial investigators. Considering the demands of judicial practice, we designated the extracted entities into the following 10 fine-grained categories: *Suspect*, *Victim*, *Stolen Item*, *Criminal Tool*, *Value (of goods)*, *Stolen Currency*, *Profit of Crime*, *Organization*, *Location*, and *Time of the Case*. An example of CAILIE is demonstrated in Fig. 1. The distribution of the dataset is presented in Table 1.

Suspect and *Victim* refer to the names, nicknames, and social media accounts of the natural persons in the case text.

Stolen Item and *Criminal Tool* belong to Item, which refers to the public and private properties in the case defined in Articles 91 and 92 of the Criminal Law of the People's Republic of China. To distinguish the items accurately, the attributes representing the items (quantity, color, brand number, etc.) are also included in Item.

Value (of goods), *Stolen Currency* and *Profit of Crime* belong to Currency, which refers to legal currency recognized by national laws, including precious metal, paper, and electronic currencies. The attributes of the currency (RMB, U.S. dollar, etc.) also need to be marked to distinguish the currency types.

Time of the Case refers to the time expression for the infringement, including the calendar time (year, month, day, etc.) and non-calendar time (morning, afternoon, night, early morning, etc.).

Location refers to the geographical location information involved in the case, which should be marked as detailed as possible, including the administrative district name, street name, community name, building number, floor number, and landmark address or natural landscape. In addition, it should include location prepositions, such as "in front of the house" or "behind the building".

Organization refers to the administrative organization, enterprise organization, or non-governmental organization involved in the case.

3.1. Data collection

We collected 1500 prosecution documents of larceny charge cases released on China Judgement Online. All collected documents were anonymized. Personal information was replaced with special tokens. Anonymization was performed manually to prevent personal information from appearing in the dataset.

3.2. Annotation

Formally, our annotation procedure included the following steps:

(1) Divide the collected legal documents into sentences;

(2) Use PYLTP1 from SCIR-HIT to extract entities from the sentences;

(3) Divide the sentences into five groups and assign them to five groups of annotators, annotating with the BIO annotation strategy;

(4) Exchange annotated data for secondary checks;

(5) Code in Python to replace privacy information with special tags; and

(6) Check private information that is not anonymized.

Ten annotators comprising undergraduate students majoring in computer science and technology participated in the task. They were divided into five groups and given five sets of data, respectively. The annotators in the same group annotated the same data but worked independently. After the annotation was completed, they conducted a consistency check of the annotated sub-datasets and achieved unanimity after discussion within the group. After the intra-group annotation was completed, the annotators cross-checked the data from different groups (each group checked the data annotated by the other group). As shown in Table 1, there are 27,031 entities in CAILIE. The distribution of entity categories in CAILIE is unbalanced. Suspect and Stolen Item account for 23.91% and 21.77% of the dataset, respectively, while Profit of Crime accounts for only 1.78%. The full dataset is divided into training, development, and test sets using the BIO annotation strategy. Statistics about sentences and tokens are shown in Table 2.

Table 1. Distribution of fine-grained (f) and coarse-grained (c) classes in the dataset.

Classes			**#**	**%**
f	Suspect	NHCS	6463	23.91
f	Victim	NHVI	3108	11.50
c	**Person**	–	**9571**	**35.41**
f	Stolen item	NASI	5884	21.77
f	Criminal tool	NATS	731	2.70
c	**Item**	–	**6615**	**24.47**
f	Value of goods	NCGV	2090	7.73
f	Stolen currency	NCSM	915	3.39
f	Profit of crime	NCSP	481	1.78
c	**Currency**	–	**3486**	**12.90**
f	Time of the case	NT	2765	10.23
c	**Time**	–	**2765**	**10.23**

Exhibit 1026 Page 202

Classes			#	%
f	Location	NS	3815	14.11
c	**Location**	–	**3815**	**14.11**
f	Organization	NO	779	2.88
c	**Organization**	–	**779**	**2.88**
Total			**27031**	**100**

Table 2. Sentence and token statistics. The word “Annotated” means annotated token.

	Sentence	Token	Annotated
Train	4000	258507	145396
Dev	627	40895	22972
Test	620	40855	23990
Total	5247	340257	192358

4. Experiments

To understand the characteristics of this dataset more profoundly, we conducted a series of experiments and analyzed the performance of the model on different labels.

Table 3. Precision, recall and F1 values of non-pretrained models on CAILIE.

	CRF			RNN+CRF			LSTM+CRF		
	Prec %	**Rec %**	**F1%**	**Prec %**	**Rec %**	**F1%**	**Prec %**	**Rec %**	**F1%**
Suspect	59.20	71.13	64.62	89.42	91.54	90.47	91.77	91.47	91.62
Victim	51.07	33.59	40.53	88.64	86.56	87.59	91.05	89.06	90.05
Stolen item	49.31	44.06	46.54	61.28	69.28	65.03	70.65	70.65	70.65
Criminal tool	57.14	16.84	26.02	62.71	58.42	60.49	70.34	53.68	60.90
Value of goods	75.08	94.63	83.73	94.93	96.82	95.87	95.82	95.63	95.72
Stolen currency	67.37	60.66	63.84	77.56	75.36	76.44	80.29	79.15	79.71
Profit of crime	0.00	0.00	0.00	69.23	91.01	78.64	67.77	92.13	78.10
Time of the case	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

	CRF			RNN CRF			LSTM CRF		
	Prec %	Rec %	F1%	Prec %	Rec %	F1%	Prec %	Rec %	F1%
Location	50.60	50.22	50.41	66.91	70.30	68.56	76.74	68.03	72.12
Organization	34.29	32.26	33.24	66.36	78.49	71.92	70.00	82.80	75.86
	56.11	48.67	52.12/40.89	76.30	72.02	74.1/69.50	82.19	72.26	76.9/71.49

Table 4. Precision, recall and F1 values of pretrained models on CAILIE.

	BERT			BERT+CRF		
	Prec%	Rec%	F1% (/MaF)	Prec%	Rec%	F1% (/MaF)
Suspect	95.56	97.47	96.51	96.12	97.20	96.66
Victim	92.75	94.75	93.74	92.74	96.61	94.63
Stolen item	77.46	81.52	79.44	78.32	81.97	80.10
Criminal tool	78.39	85.25	81.68	75.12	84.15	79.38
Value of goods	97.21	96.63	96.92	97.20	96.24	96.72
Stolen currency	84.26	83.42	83.84	88.08	85.43	86.74
Profit of crime	77.23	95.12	85.25	71.96	93.90	81.48
Time of the case	93.47	91.52	92.48	93.33	91.67	92.49
Location	81.20	81.20	81.20	80.92	82.92	81.91
Stolen currency organization	80.84	91.53	85.86	82.63	93.12	87.56
	87.43	89.61	88.51/87.69	87.64	90.13	88.87/87.77

4.1. Baseline models

We conducted experiments on the following models: (1) Conditional Random Fields (CRFs): CRFs are a class of statistical modeling methods that take context into account; (2) BiLSTM + CRF: BiLSTM + CRF models contextual information by combining forward LSTM and backward LSTM; (3) Recurrent Neural Network (RNN) + CRF: RNN takes legal text data as an input sequence, and performs recursion in the evolution direction of the sequence to process variable-length sequences; (4) Bidirectional Encoder Representation from Transformers (BERT): BERT has been pretrained on a large amount of corpus, and then fine-tuned on legal texts to obtain valuable features from legal documents; (5) BERT + CRF: BERT provides powerful feature extraction capabilities, and CRF makes the model more suitable for sequence labeling problems.

To measure the difficulty of extracting different kinds of information, we calculated the F1 score of the extraction results of the model in each type of entity. Considering the imbalance of the data distribution, both Macro F1 and Micro F1 scores were computed.

4.2. Results

We chose four metrics to compare the performance of the widely used models, including micro-precision, micro-recall, micro-F1, and macro-F1. The main results of the non-pretrained and pretrained models are shown in Table 3, Table 4, respectively. The pretrained models achieve better performance in recognizing all kinds of entities. As expected, BERT + CRF performs best (see Table4). Compared with BERT, BERT + CRF improved the F1-scores of stolen currency prediction and organization prediction by 2.90% and 1.70%, respectively. Overall, the models perform comparably to BERT. The performance remains mostly unchanged when CRF was dropped, indicating that BERT can obtain almost all the features provided by CRF.

From another point of view, BERT shows a huge leap in performance over LSTM + CRF (+ 11.97 Micro F1) and RNN + CRF (+ 14.77 Micro F1). Meanwhile, the performance of the non-pretrained models on Item (Stolen Items and Criminal Tools) prediction and Time of the Case prediction is much worse than that of the pretrained models. The F1 score for Time of the Case reaches 92.49 and 79.38 for Criminal Tools. Both pretrained models perform well in extracting the Time of the Case and Items (Stolen items and Criminal tools), which are difficult for non-pretrained models (see Table3). This is due to the fact that large-scale pre-training provides additional knowledge, and the correlation between tokens can be directly calculated without passing through hidden layers. In the face of highly diverse entities (e.g., Criminal Tools, Organizations) and long sequence entities (e.g., location), the pre-trained model is capable of identifying them accurately. The pre-trained language model is also more capable of identifying confusing entities. Confusing entities, such as Stolen Currency and Profit of Crime, usually belong to the same coarse-grained entity category, and their expressions are quite similar. The pre-trained models improve the F1 score of Stolen Currency by about 6 percentage points. However, we see significant room for model improvement on this task.

5. Overview of CAIL challenge

CAILIE was used for the competition of the CAIL2021 information extraction track. The competition set up two tracks, namely, deep model and non-deep model, for this dataset. A total of 75 teams participated in the track in CAIL2021. Only one team participated in the non-deep model track. The performance of the submitted model was far from the deep models, which indicates the powerful ability of deep learning models to solve the task of information extraction. Therefore, we only analyzed the solution strategy of the deep learning model. Based on the models submitted by the entrants, we provide the following tips for improving the performance of the models:

(1) Entity length limitation. Some models limit the maximum length of short entities such as Suspect and Victim. When the model extracts nested entities, the same entity is occasionally

identified as nested entities of different lengths, and the model selects the relatively long or short one.

(2) Pointer network. The pointer network converts the sequence labeling task into the prediction task of the start and end positions of entities. This method of first blocking and then classifying is very suitable for nested NER. Therefore, a considerable number of entrants adopted this method.

(3) Ensemble. Several entrants chose to train multiple models, and performed label-weighted voting at the entity level. Only the entities predicted by the models that are greater than a threshold number can be selected.

6. Future direction

Given the focus on sentencing standards in this dataset, CAILIE has the potential to be used for the downstream tasks of legal artificial intelligence. Similar case matching is the process of finding cases with the same criminal circumstances as the specified cases from the case database to support automatic sentencing, legal recommendation, and other intelligent auxiliary case handling. Therefore, this dataset provides important matching support. Statutory reasoning, legal argument mining, and legal judgment prediction provide approximate requirements. The presentation of CAILIE is valuable in providing legal-level interpretability to the model. Nonetheless, a number of problems still need to be solved.

Boundary Detection. Inaccurate entity boundary recognition is more common in some entities with large differences in length and insignificant common features of the boundary. Since the Chinese do not divide words, it is difficult to identify entity boundaries when information extraction is performed as a sequence labeling task, such as the front boundaries of Location and Organization.

Confusing Labels. In legal information extraction, the similarity between certain entity types is high. For example, Value(of goods), Stolen Currency, and Profit of Crime all belong to Currency, and only a few of the entities representing time belong to Time of the Case. The representations of these different entities are similar, so it is necessary to obtain information from the context to distinguish them accurately.

Nested Entities. Nested entity recognition has always been a primary challenge for named entity recognition. Due to the characteristics of judicial documents, the proportion of nested entities in this dataset is quite large, such as the most frequently occurring nests, namely, victims and stolen items. The effective identification of nested entities will assist judicial case investigators in quickly reading files and clarifying case information. In addition, it is an important basis for a series of tasks such as knowledge graph construction, similar case recommendation, and automatic sentencing recommendation. Therefore, we are keen on improving the accuracy of nested entity recognition.

7. Conclusion

In this paper, we propose a Chinese theft crime information extraction dataset called CAILIE, which contains more than 27,000 entities in 10categories. Compared with existing datasets, CAILIE focuses on fine-grained entities unique to legal documents and views them from the perspective of prosecutors. The experimental results show that Chinese legal information extraction technology still has room for improvement. In the future, we will try to enrich CAILIE with relation information and explore more ways to assist judicial case handling.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2022YFC3301801), the Fundamental Research Funds for the Central Universities, China (No. DUT22ZD205).

Ethics statement

As part of the dataset creation process, we collected original data from publicly accessible legal documents online. In the data-processing stage, we attached importance to the ethical decisions of the regulators, strictly abided by the ethical norms, and anonymized all private information in the dataset. Nonetheless, we acknowledge that there is still a risk of information leakage during the data dissemination process. We will work hard to ensure data security, and strictly check the quality of anonymization of private information when releasing data.

Recommended articles

References

Aumiller et al., 2022 Aumiller Dennis, Chouhan Ashish, Gertz Michael
EUR-lex-sum: A multi- and cross-lingual dataset for long-form summarization in the legal domain
(2022)
Google Scholar ↗

Bhattacharya et al., 2021 Bhattacharya Paheli, Poddar Soham, Rudra Koustav, Ghosh Kripabandhu, Ghosh Saptarshi

Incorporating domain knowledge for extractive summarization of legal case documents

(2021)

CoRR abs/2106.15876 ↗

Google Scholar ↗

Bibal et al., 2020 Bibal Adrien, Lognoul Michael, Streel Alexandre De, Frénay Benot

Legal requirements on explainability in machine learning

Artif. Intell. Law (2020)

Google Scholar ↗

Bithel and Malagi, 2021 Bithel Shivangi, Malagi Sumitra S.

Unsupervised identification of relevant prior cases

(2021)

CoRR abs/2107.08973 ↗

Google Scholar ↗

Cardellino et al., 2017 Cardellino C., Teruel M., Alemany L.A., Villata S.

A low-cost, high-coverage legal named entity recognizer, classifier and linker

International Conference on Artificial Intelligence and Law (2017)

Google Scholar ↗

Chen et al., 2020 Chen Y., Sun Y., Yang Z., Lin H.

Joint entity and relation extraction for legal documents with legal feature enhancement

International Conference on Computational Linguistics (2020)

Google Scholar ↗

Do et al., 2017 Do Phong-Khac, Nguyen Huy-Tien, Tran Chien-Xuan, Nguyen Minh-Tien, Nguyen Minh-Le

Legal question answering using ranking SVM and deep convolutional neural network

(2017)

CoRR abs/1703.05320 ↗

Google Scholar ↗

Dozier et al., 2010 Dozier C., Kondadadi R., Light M., Vachher A., Wudali R.

Named Entity Recognition and Resolution in Legal Text

DBLP (2010)

Google Scholar ↗

Duan et al., 2019a Duan Xingyi, Wang Baoxin, Wang Ziyue, Ma Wentao, Cui Yiming, Wu Dayong, Wang Shijin, Liu Ting, Huo Tianxiang, Hu Zhen, Wang Heng, Liu Zhiyuan

CJRC: A reliable human-annotated benchmark DataSet for Chinese judicial reading comprehension

(2019)

CoRR abs/1912.09156 ↗

Google Scholar ↗

Duan et al., 2019b Duan,Xinyu, Zhang,Yating, Yuan,Lin, Zhou,Xin, Liu,Xiaozhong, Wang,Tianyi, Wang,Ruocheng, Zhang,Qiong, Sun,Changlong, Wu,Fei, 2019b. Legal summarization for multi-role debate dialogue via controversy focus mining and multi-task learning. In: Proceedings of CIKM. pp. 1361–1370.

Google Scholar ↗

Gil Semo et al., 2022 Gil Semo Dor Bernsohn, Hagag Ben, Hayat Gila, Niklaus Joel

ClassActionPrediction: A challenging benchmark for legal judgment prediction of class action cases in the US

(2022)

Google Scholar ↗

Gong et al., 2019 Gong Q., Cheng Y., Chen J., Chao L.I., Zhang D., Long Y.

Research on the recognition model of court judgment named entity based on deep learning

J. Hubei Univ. Technol. (2019)

Google Scholar ↗

Habernal et al., 2022 Habernal Ivan, Faber Daniel, Recchia Nicola, Bretthauer Sebastian, Gurevych Iryna

Mining legal arguments in court decisions

(2022)

Google Scholar ↗

Holzenberger et al., 2020 Holzenberger Nils, Blair-Stanek Andrew, Durme Benjamin Van

A dataset for statutory reasoning in tax law entailment and question answering

(2020)

CoRR abs/2005.05257 ↗

Google Scholar ↗

Huang et al., 2021 Huang Yunyun, Shen Xiaoyu, Li Chuanyi, Ge Jidong, Luo Bin

Dependency learning for legal judgment prediction with a unified text-to-text transformer

(2021)

CoRR abs/2112.06370 ↗

Google Scholar ↗

Iftikhar et al., 2019 Iftikhar A., Jaffry Swuq, Malik M.K.

Information mining from criminal judgments of lahore high court

IEEE Access (2019)

Google Scholar ↗

Jackson et al., 2003 Jackson P., Al-Kofahi K., Tyrrell A., Vachher A.

Information extraction from case law and retrieval of prior cases

Artificial Intelligence, 150 (1–2) (2003), pp. 239-290

View PDF View article View in Scopus ↗ Google Scholar ↗

Kanapala et al., 2019 Kanapala Ambedkar, Pal Sukomal, Pamula Rajendra

Text summarization from legal documents: a survey

Artif. Intell. Rev., 51 (3) (2019), pp. 371-402

Crossref ↗ View in Scopus ↗ Google Scholar ↗

Kien et al., 2020 Kien, Phi Manh, Nguyen, Ha-Thanh, Bach, Ngo Xuan, Tran, Vu, Le Nguyen, Minh, Phuong, Tu Minh, 2020. Answering Legal Questions by Learning Neural Attentive Text Representation. In: Proceedings of COLING. pp. 988–998.

Google Scholar ↗

Kuru et al., 2016 Kuru, Onur, Can, Ozan Arkan, Yuret, Deniz, 2016. CharNER: Character-Level Named Entity Recognition. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016. pp. 911–921.

Google Scholar ↗

Kwak et al., 2022 Kwak Alice Saebom, Israelsen Jacob O., Morrison Clayton T.

Validity assessment of legal will statements as natural language inference

(2022)

Google Scholar ↗

Li et al., 2021 Li Jingye, Fei Hao, Liu Jiang, Wu Shengqiong, Zhang Meishan, Teng Chong, Ji Donghong, Li Fei

Unified named entity recognition as word-word relation classification

(2021)

CoRR abs/2112.10070 ↗

Google Scholar ↗

Luoma and Pyysalo, 2020 Luoma Jouni, Pyysalo Sampo

Exploring cross-sentence contexts for named entity recognition with BERT

(2020)

CoRR abs/2006.01563 ↗

Google Scholar ↗

Ma et al., 2021 Ma Luyao, Zhang Yating, Wang Tianyi, Liu Xiaozhong, Ye Wei, Sun Changlong, Zhang Shikun

Legal judgment prediction with multi-stage CaseRepresentation learning in the real court setting

(2021)
CoRR abs/2107.05192 ↗

Google Scholar ↗

Mandal et al., 2017 Mandal A., Chaki R., Saha S., Ghosh K., Ghosh S.
Measuring similarity among legal court case documents
The 10th Annual ACM India Compute Conference (2017)

Google Scholar ↗

Nguyen et al., 2021 Nguyen Duy-Hung, Nguyen Bao-Sinh, Nghiem Nguyen-Viet-Dung, Le Dung Tien, Khatun Mim Amina, Nguyen Minh-Tien, Le Hung
Robust deep reinforcement learning for extractive legal summarization
(2021)
CoRR abs/2111.07158 ↗

Google Scholar ↗

Nguyen et al., 2020 Nguyen Ha-Thanh, Vuong Hai-Yen Thi, Nguyen Phuong Minh, Dang Tran Binh, Bui Quan Minh, Sinh Vu Trong, Nguyen Chau Minh, Tran Vu D., Satoh Ken, Nguyen Minh Le
JNLP team: Deep learning for legal processing in COLIEE 2020
(2020)
CoRR abs/2011.08071 ↗

Google Scholar ↗

Nie et al., 2020 Nie,Y., Tian,Y., Wan,X., Song,Y., Dai,B., 2020. Named Entity Recognition for Social Media Texts with Semantic Augmentation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Google Scholar ↗

Niklaus et al., 2021 Niklaus Joel, Chalkidis Ilias, Stürmer Matthias
Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark
(2021)
CoRR abs/2110.00806 ↗

Google Scholar ↗

Shao et al., 2021 Shao Y., Wu Y., Liu Y., Mao J., Ma S.
Investigating user behavior in legal case retrieval
SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (2021)

Google Scholar ↗

Shen et al., 2022 Shen Zejiang, Kyle Lo Lauren Yu, Dahlberg Nathan, Schlanger Margo, Downey Doug
Multi-LexSum: Real-world summaries of civil rights lawsuits at multiple granularities
(2022)

Google Scholar ↗

Skylaki et al., 2020 Skylaki Stavroula, Oskooei Ali, Bari Omar, Herger Nadja, Kriegman Zac

Named entity recognition in the legal domain using a pointer generator network

(2020)

CoRR abs/2012.09936 ↗

Google Scholar ↗

Tran et al., 2020 Tran Vu D., Nguyen Minh Le, Satoh Ken

Building legal case retrieval systems with lexical matching and summarization using A pre-trained phrase scoring model

(2020)

CoRR abs/2009.14083 ↗

Google Scholar ↗

Wang and Jin, 2020 Wang C., Jin X.

Study on the multi-task model for legal judgment prediction

2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA) (2020)

Google Scholar ↗

Wang et al., 2022 Wang Shuhe, Li Xiaoya, Meng Yuxian, Zhang Tianwei, Ouyang Rongbin, Li Jiwei, Wang Guoyin

k NN-NER: Named entity recognition with nearest neighbor search

(2022)

arXiv preprint arXiv:2203.17103 ↗

Google Scholar ↗

Wang et al., 2021 Wang Zihan, Song Hongye, Ren Zhaochun, Ren Pengjie, Chen Zhumin, Liu Xiaozhong, Li Hongsong, de Rijke Maarten

Cross-domain contract element extraction with a bi-directional feedback clause-element relation network

Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM (2021)

Google Scholar ↗

Wehnert et al., 2019 Wehnert Sabine, Hoque Sayed Anisul, Fenske Wolfram, Saake Gunter

Threshold-based retrieval and textual entailment detection on legal bar exam questions

(2019)

CoRR abs/1905.13350 ↗

Google Scholar ↗

Westermann et al., 2021 Westermann Hannes, Savelka Jaromir, Benyekhlef Karim

Paragraph similarity scoring and fine-tuned BERT for legal information retrieval and entailment

New Frontiers in Artificial Intelligence (2021)

Google Scholar ↗

Xiao et al., 2021 Xiao Chaojun, Hu Xueyu, Liu Zhiyuan, Tu Cunchao, Sun Maosong

Lawformer: A pre-trained language model for Chinese legal long documents

(2021)

CoRR abs/2105.03887 ↗

Google Scholar ↗

Yang et al., 2019 Yang W., Jia W., Zhou X., Luo Y.

Legal judgment prediction via multi-perspective bi-feedback network

Twenty-Eighth International Joint Conference on Artificial Intelligence IJCAI-19 (2019)

Google Scholar ↗

Yin et al., 2018 Yin Xiaoxiao, Zheng Daqi, Lu Zhengdong, Liu Ruifang

Neural entity reasoner for global consistency in NER

(2018)

CoRR abs/1810.00347 ↗

Google Scholar ↗

Zeng et al., 2005 Zeng Y., Wang R., Zeleznikow J., Kemp E.

Knowledge representation for the intelligent legal case retrieval

International Conference on Knowledge-Based & Intelligent Information & Engineering Systems (2005)

Google Scholar ↗

Zhang et al., 2017 Zhang N., Pu Y.F., Yang S.Q., Zhou J.L.

An ontological Chinese legal consultation system

IEEE Access, PP (99) (2017), p. 1

View PDF View article Google Scholar ↗

Zhong et al., 2020 Zhong,Haoxi, Xiao,Chaojun, Tu,Cunchao, Zhang,Tianyang, Liu,Zhiyuan, Sun,Maosong, 2020. Jec-qa: A legal-domain question answering dataset. In: Proceedings of the AAAI.

Google Scholar ↗

Zhou et al., 2022 Zhou Kang, Li Yuepei, Li Qi

Distantly supervised named entity recognition via confidence-based multi-class positive and unlabeled learning

(2022)

Google Scholar ↗

Zhou et al., 2019 Zhou Yi, Zheng Xiaoqing, Huang Xuanjing

Chinese named entity recognition augmented with lexicon memory

(2019)

CoRR abs/1912.08282 ↗

Google Scholar ↗

Cited by (7)

Legal knowledge infusion for large language models: A survey

2026, Information Fusion

Show abstract

Graph-Enhanced Prompt Learning for Cross-Domain Contract Element Extraction ↗

2025, ACM Transactions on Information Systems

LAiW: A Chinese Legal Large Language Models Benchmark ↗

2025, Proceedings International Conference on Computational Linguistics Coling

LegalATLE: an active transfer learning framework for legal triple extraction ↗

2024, Applied Intelligence

Judicial named entity recognition enhanced with semantic and boundary ↗

2024, Qinghua Daxue Xuebao Journal of Tsinghua University

Information Extraction from Legal Wills: How Well Does GPT-4 Do? ↗

2023, Findings of the Association for Computational Linguistics Emnlp 2023

View all citing articles on Scopus ↗

© 2022 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd.



All content on this site: Copyright © 2026 Elsevier B.V., its licensors, and contributors. All rights are reserved, including those for text and data mining, AI training, and similar technologies. For all open access content, the relevant licensing terms apply.

APPENDIX 1011J

The ACL Anthology: Current State and Future Directions

Daniel Gildea
Department of Computer Science
University of Rochester
gildea@cs.rochester.edu

Min-Yen Kan
School of Computing
National University of Singapore
kanmy@comp.nus.edu.sg

Nitin Madnani
Educational Testing Service
Princeton, NJ
nmadnani@ets.org

Christoph Teichmann and **Martín Villalba**
Department of Language Science and Technology
Saarland University
villalba@coli.uni-saarland.de

Abstract

The Association of Computational Linguistic's Anthology is the open source archive, and the main source for computational linguistics and natural language processing's scientific literature. The ACL Anthology is currently maintained exclusively by community volunteers and has to be available and up-to-date at all times. We first discuss the current, open source approach used to achieve this, and then discuss how the planned use of Docker images will improve the Anthology's long-term stability. This change will make it easier for researchers to utilize Anthology data for experimentation. We believe the ACL community can directly benefit from the extension-friendly architecture of the Anthology. We end by issuing an open challenge of reviewer matching we encourage the community to rally towards.

1 Introduction

The ACL Anthology[1] is a service offered by the Association for Computational Linguistics (ACL) allowing open access to the proceedings of all ACL sponsored conferences and journal articles. As a community goodwill gesture, it also hosts third-party computational linguistics literature from sister organizations and their national venues. It offers both text and faceted search of the indexed papers, author-specific pages, and can incorporate third-party metadata and services that can be embedded within pages (Bysani and Kan, 2012). As of this paper, it hosts over 43,000 computational linguistics and natural language processing papers, along with their metadata. Over 4,500 daily requests are served by the Anthology. The code for the Anthology is available at https://github.com/acl-org/acl-anthology under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 International License[2]. Slightly different from the Anthology source code, ACL also licenses its papers with a more liberal license, supporting Creative Commons Attribution 4.0 International License[3], supporting liberal re-use of papers published with the ACL.

The maintenance of the code and the website is handled through volunteer efforts coordinated by the Anthology editor. Running a key service for the computational linguistics community that needs to be continuously available and updated frequently is one of the main issues in administering the Anthology.

We discuss this issue along with the challenges of running a large scale project on a volunteer basis and its resulting technical debt. As we look towards the future, previous research has shown that it can also be used as a data source to characterize the work and workings of the ACL community (Bird et al., 2008; Vogel and Jurafsky, 2012; Anderson et al., 2012). Extensions to the Anthology that build on this information could make the Anthology an even more valuable resource for the community. We will discuss two possibl eextensions – anonymous pre-prints and support for finding relevant submission reviewers by linking au-

[1] https://aclanthology.info/

[2] https://creativecommons.org/licenses/by-nc-sa/3.0/

[3] https://creativecommons.org/licenses/by/4.0/

Proceedings of Workshop for NLP Open Source Software, pages 23–28
Melbourne, Australia, July 20, 2018 (c) 2018 Association for Computational Linguistics
Exhibit 1028 Page 217

thors in the Anthology with their research interests and community connections. Beyond being useful in itself, work on such challenges has the potential to motivate the ACL community to further support the Anthology.

2 Current State of the Anthology

The ACL Anthology was proposed as a project to the ACL Executive by Steven Bird at the 2001 ACL conference and first launched in 2002, with a second version developed in 2012, commissioned by the ACL committee. Steven Bird also served as the first editor of the anthology from 2002 to 2007, a post which Min-Yen Kan took over in 2008 and continues to fill as of today. The Anthology provides access to papers in Portable Document Format (PDF) as well as the associated metadata in multiple formats (e.g., BIBTEX and Endnote). For recent papers, authors can also opt include data, notes and open-source software, and may provide Digital Object Identifiers (DOIs) for permalinking the citations within their papers.

The technology behind the current version is detailed in Table 1. As a community project, daily administration and development is handled by volunteers. However, to tackle larger problems with the Anthology which require a more focused effort, the ACL committee has solicited paid assistance. Hosting and bandwidth for the Anthology has historically been provided by universities free of charge. It was hosted at the National University of Singapore until the spring of 2017, when it was migrated to its current home at Saarland University. In the future, hosting duties are planned to fall under the umbrella of the ACL itself, unifying all services under https://www.aclweb.org/portal/.

Framework	Ruby on Rails
Search engine	Solr
Database	PostgreSQL
Web server (Prod./Test)	Nginx / Jetty
Operating System	Debian GNU-Linux

Table 1: Tech stack for the ACL Anthology.

The most important task is the importing, indexing and provisioning of newly accepted papers from recent conference proceedings and journal issues. The original Anthology defined an XML format for simple bibliographic metadata, which has been extended to support the more recent features of associated software, posters, videos and datasets that accompany the scholarly publications. Providing the XML for new materials is an semi-automated process that is largely integrated with the various mechanisms for managing ACL conference submissions and printed proceedings. It is straightforward for ACL events that utilize the licensed START conference management software[4], as an established software pipeline builds upon the artefacts used for creation of the final publications themselves. After the accepted papers are finalized, START produces an archive file of camera-ready PDF files and author-provided metadata such as the title, author list, and abstract for each paper. These files are processed by a set of scripts in START maintained by ACL publication chairs in order to assign page numbers to papers, and to produce a PDF proceedings volume for each conference complete with a table of contents, author index, and other front matter. These scripts also produce bibliographic information that are programmatically transformed into the ACL Anthology's XML format. The Anthology is then updated with the author-provided PDFs and the XML metadata. For importing journal articles and venues not using the START submission system, additional manual work is necessary to construct the Anthology XML. Sanity checks and some manual curation is also necessary to deal with issues such as character encodings and accents in names, multipart family names, and so on. This pipeline has reached a point of high efficiency, but may need to be adapted if the ACL ever considers it necessary to integrate with a different service for conference organization.

3 Running the Anthology as a Community Project

Since the Anthology is not tied to a specific research project or institution, contributors that work on Anthology-related system administration and development tasks have been recruited in response to calls for volunteers at the main ACL conferences. In contrast, new features have been developed by researchers using the ACL Anthology as a resource in their own work, unconnected with the daily operation of the Anthology. Such research deliverables include, for example, the creation of a corpus of research papers (Bird et al., 2008), an author citation network (Radev et al., 2013) or a

[4] https://www.softconf.com/

Exhibit 1026 Page 218

faceted search engine (Schäfer et al., 2012; Buitelaar et al., 2014). These factors, in combination with the multiple, changing responsibilities and shifting research interests of community members, mean that new volunteers join and leave the Anthology team in unpredictable and sporadic patterns. Preserving knowledge about the Anthology's operational workflow is thus one of the most important challenges for the Anthology.

The Anthology editor has played a key role ensuring the continuity of the entire project. This position has so far always been filled for multiple years, longer than the normal time frame for an ACL officer. The role has been critical in ensuring a smooth transition between volunteers, at the cost of a long term with a heavy workload and a potential single point of failure. In order to tackle both issues, there is currently a concerted effort to improve the documentation of all tasks related to maintaining the Anthology.

As the ACL community and its publishing needs continue to grow, the ACL Executive is considering commercial support for publishing. While this may be suitable for help with daily operations, we strongly advocate the continuation and promotion of a closely-knit volunteer group for development. Passing the responsibilities for the Anthology to a commercial devoid who has no intrinsic interest in the Anthology's scientific contents may end up poorly.

4 Future Proofing the Anthology

All code, documentation, bug reports, and feature requests are hosted at `https://github.com/acl-org/acl-anthology`, along with instructions detailing the steps required to set up an instance of the Anthology and keep it updated with proceedings for new conferences. These instructions have been verified and updated using test builds. We began with the initial documentation provided by experienced contributors to the project and the original developer. New volunteers were then asked to set up and update a new instance of the Anthology on a new server while communicating with more experienced contributors. The documentation was expanded and updated based on the problems and questions encountered during this process. The resulting documentation will likely reduce the learning curve for new volunteers and will make their recruitment easier. It will also make it easier to migrate the Anthology to new servers when the hosting arrangement changes or to create mirrors. The latter is an important future task for the Anthology in order to ensure that alternatives are available if the main Anthology server experiences any downtime.

The current implementation of the Anthology has been extended over the years with minor improvements to functionality and bug fixes. The core code has remained mostly intact from its original version and has proved to be robust and reliable. However, fearing the introduction of bugs and instability (Spolsky, 2000), the maintainers chose to keep the software working in its current state for as long as the technology would allow it, and focus their resources instead on features that would help the community with their research and publication efforts.

This choice is not without its drawbacks. One key problem is the deprecation of dependencies with time. For example, Ruby 2.0 is no longer available in Debian repositories, and SSL support no longer compiles against it by default. These problems can be seen as indicators that delaying upgrades might not be feasible for much longer. Where possible, deprecated libraries are replaced with newer versions. This is the case for the database, web server, and the Java interpreter, all of which have been replaced with little extra effort. When a new version of a library breaks backwards compatibility, the software is either upgraded or frozen in its current version. Ruby (frozen at `2.0.0-p353` via RVM) and Solr are both examples of the latter, with detailed documented instructions to replicate the software environment.

In addition to the production Anthology site, a second version is kept on low-cost cloud servers for testing purposes. This copy has proven useful for testing step-by-step instructions, since rolling back the server to a clean state requires neither authorization nor downtime. It is also used as a staging area, and to do trial imports of new proceedings and for volunteer training.

Security is another major concern: older dependencies increase exposure to unpatched bugs. The Anthology currently does not collect or store personal data, rendering the consequences of a data breach modest. A compromised server, however, presents not only a risk for the maintainers (service downtime, unauthorized applications) but for the community at large, due to the large number of researchers who could be exposed to malicious

Exhibit 1026 Page 219

scripts. While the former puts the goodwill of the hosting institution at risk, the latter would affect a large portion of the ACL community.

To tackle issues with outdated software, the Anthology volunteer group is working on making the entire Anthology available via a Docker image (Matthias and Kane, 2015). Docker provides a virtualized environment (also known as a *container*) in which software can be run but where, unlike a virtual machine, the underlying operating system resources can be used directly. Containers are typically stateless, allowing system administrators to add and restart services with minimum friction. Hosting a mirror of the Anthology with Docker containers abstracts away the relatively complex server setup and makes it easier to tackle dependency problems independently from future mirror deployments. As a result, hosting institutions can apply their own internal security policies, and the community can benefit from the added robustness via a larger network of mirrors. Development versions of this image are already available at `https://github.com/acl-org/anthology-docker`. When an instance of this Docker container is started, it first downloads all the data necessary to run the Anthology, inclusive of the metadata and source publications (PDF files) for all proceedings hosted within the Anthology. The resulting Anthology instance is a peer of the production site, but completely independent. This makes it possible for member institutions and even interested individual members to easily provide a mirror or experiment with the data in the Anthology.

Freezing software versions has proven useful to keep stability under control, improve documentation practices, and implement long-requested features like search engine indexing. This does not preclude a full software upgrade from being part of our development roadmap. With better test coverage and expanded consistency checks in place, we expect the first successful upgrade tests to be within our reach in the near future.

Docker containers and temporary servers also show great promise for researchers. An isolated, easy-to-replicate software environment reduces friction in transferring tools between researchers usually caused by incompatible software, simplifies the replication of experiments, and limits the data loss due to software bugs. A container-like approach specifying complete environments can also help in distributing code and general research within the community (e.g., CodaLab[5] as used in SemEval competitions). In the future, best practices within the community may encourage researchers to program and experiment within Docker images to aid reproducibility.

The Anthology is currently stable and supports its current, intended use. However, to ensure that the ACL Anthology continues fulfilling its key roles, we call on the members of the ACL to help with both its operational and development goals:

- hosting mirrors of the Anthology and developing policy for mirror management;
- adding and indexing new publications to the Anthology;
- maintaining and updating the code underlying the Anthology;
- extending the capabilities of the Anthology to help tackle new challenges facing the ACL.

5 Challenges for the Anthology

Maintaining community buy-in for the Anthology is necessary to ensure its future. This is best assured by extending the Anthology with useful capabilities that align with research efforts. This is crucially enabled by the liberal licensing scheme that the ACL employs for the publications to empower end users. Research on the history and structure of the NLP community based on this data has already been undertaken (Anderson et al., 2012; Vogel and Jurafsky, 2012).

Anonymous Pre-prints. A current challenge needing attention is the result of the increasing popularity of pre-prints and their role in promoting scientific progress. However, such pre-print systems are not anonymous, interfering with the well-documented gains that author-blinded publications help in combating bias. Through membership polls and subcommitee study, the ACL executive has adopted a recent set of guidelines upholding the value of double-blinded submissions (ACL Executive Committee, 2017).

One solution would be the use of anonymous pre-prints as an option for authors. Currently two ways of implementing this have been discussed: as a collaboration with an existing pre-print service such as arXiv[6] or through hosting pre-prints

[5] `https://worksheets.codalab.org/`
[6] `https://arxiv.org/`

Exhibit 1026 Page 220

directly within the Anthology. While the latter option would be a challenge to the Anthology – requiring increased resources both for monitoring the submissions and for scaling the system architecture to a larger and less controlled inflow of papers – but could result in better community control of the process, and a greater awareness and feeling of co-ownership of the Anthology and its data among ACL members.

Reviewer Matching. One key problem with scientific conference and journal organization is in finding suitable reviewers for the peer review process, which is also a key problem for ACL.[7] We believe that we can leverage the ACL Anthology data to support conference organizers in the assignment of potential peer reviewers. There has been a substantial growth in the number of submissions to the main ACL conferences in recent years (Barzilay, 2017), and the ACL has been active in supporting automated approaches to solve the problem (Stent and Ji, 2018) such as the Toronto Paper Matching System (TPMS) (Charlin and Zemel, 2013). However, data for judging the fit between a reviewer and submitted papers are available in the Anthology; i.e., a reviewer's interests and expertise as encoded in their previous publications. Mining and representing such information directly from the Anthology, where data about potential reviewers is already available, makes it unnecessary to upload papers to an external platform, mitigating current low response rates. Measuring overlap between reviewer interests and a submitted paper, based on the reviewer's previous publications, is a problem that the NLP community is ideally suited to solve. Furthermore, the information generated by such a tool could serve conference chairs and journal editors when considering how much weight to assign to a review from specific reviewers. The data required for building such a tool would be both the text and metadata from every submitted paper. While some metadata is already accessible within the Anthology, clean textual content of papers would need to be harvested from the source PDF files, which currently has been partially achieved. (Bird et al., 2008) suggests that the text can generally be extracted using standard tools, with additional processing only necessary for a small fraction of the data. We are aware that clean textual data from the Anthology archives is current on-going interest being investigated by a number of NLP/CL teams within the community.

If such a solution were to be implemented, it would be in the interest of the entire community to have the Anthology maintainers integrate it directly into the Anthology, with support from the original implementers. This has been a problem in the past, where attempts to extend the capabilities of the Anthology with more detailed search and annotation (Schäfer et al., 2011, 2012) were spun off as independent systems to start with and have still not become part of the Anthology service.

We note that these two challenges are synergistically solved. Solving the first challenge will provide the submissions' source text within the Anthology framework and promote better coupling for the second challenge of reviewer matching.

6 Conclusion

The ACL Anthology is a key resource for researchers in the NLP community. We have described the software engineering and maintenance work that goes on behind-the-scenes in order for the Anthology to serve its purpose. This includes ingestion of new papers, maintenance of the Anthology codebase, and the social aspects of recruiting volunteers for this work. The task of training future volunteers and ensuring Anthology uptime is likely to become easier due to improved documentation and simplified server set-up. However, recruitment of new volunteers continues to be an issue.

We invite all community members to download the Anthology images for experimentation, not only for the challenge of automated reviewer assignment, but also for other use cases based on their own research interests. We hope that open challenges and the tasks associated with extending the usefulness of the Anthology will motivate more community members to take interest and become and familiar with its inner workings. We extend an open invitation to anyone interested in the Anthology to get in touch with the members of the team. Our current needs are focused on system administration, software development, database management, and Docker integration, but any kind of experience is welcome.

[7] As intimated through internal discussions with the ACL executive committee.

Exhibit 1026 Page 221

References

ACL Executive Committee. 2017. Acl policies for submission, review and citation. https://www.aclweb.org/adminwiki/index.php?title=ACL_Policies_for_Submission,_Review_and_Citation. [Online; accessed 05-April-2018].

Ashton Anderson, Dan McFarland, and Dan Jurafsky. 2012. Towards a Computational History of the ACL: 19802008. In *Proceedings of the ACL Special Workshop 2012 on Rediscovering 50 years of Discoveries*.

Regina Barzilay. 2017. Statistics on Submissions and Status Update. https://acl2017.wordpress.com/2017/02/15/statistics-on-submissions-and-status-update/. [Online; accessed 05-April-2018].

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Language Resources and Evaluation Conference (LREC 08)*, Marrakesh, Morocco.

Paul Buitelaar, Georgeta Bordea, and Barry Coughlan. 2014. Hot topics and schisms in NLP: Community and trend analysis with saffron on ACL and LREC proceedings. In *9th Edition of Language Resources and Evaluation Conference (LREC2014)*.

Praveen Bysani and Min-Yen Kan. 2012. Integrating User-Generated Content in the ACL Anthology. In *Proceedings of the ACL Special Workshop 2012 on Rediscovering 50 years of Discoveries*.

Laurent Charlin and Richard S. Zemel. 2013. The Toronto Paper Matching System: An automated paper-reviewer assignment system. In *ICML Workshop on Peer Reviewing and Publishing Models (PEER)*.

Karl Matthias and Sean P. Kane. 2015. *Docker: Up & Running*. O'Reilly Media, Inc.

Dragomir R. Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The ACL anthology network corpus. *Language Resources and Evaluation Journal*.

Ulrich Schäfer, Bernd Kiefer, Christian Spurk, Jörg Steffen, and Rui Wang. 2011. The ACL Anthology Searchbench. In *Proceedings of the ACL-HLT 2011 System Demonstrations*.

Ulrich Schäfer, Jonathon Read, and Stephan Oepen. 2012. The ACL Anthology Searchbench. In *Proceedings of the ACL Special Workshop 2012 on Rediscovering 50 years of Discoveries*.

Joel Spolsky. 2000. Things You Should Never Do, Part I. https://www.joelonsoftware.com/2000/04/06/things-you-should-never-do-part-i. [Online; accessed 29-March-2018].

Amanda Stent and Heng Ji. 2018. A Review of Reviewer Assignment Methods. https://naacl2018.wordpress.com/2018/01/28/a-review-of-reviewer-assignment-methods. [Online; accessed 29-March-2018].

Adam Vogel and Dan Jurafsky. 2012. He Said, She Said: Gender in the ACL Anthology. In *Proceedings of the ACL Special Workshop 2012 on Rediscovering 50 years of Discoveries*.