

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

WEALTH, INC.
Petitioner

v.

VANILLA TECHNOLOGIES, INC.
Patent Owner

Case PGR2026-_____

U.S. Patent No. 12,353,917
Title: EVENT-BASED RESOURCE ALLOCATION SYSTEM
Filing Date: 11/27/2024
Issue Date: 07/08/2025

DECLARATION OF INGRID HSIEH-YEE, PH.D.

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.
U.S. Patent No. 12,353,917

TABLE OF CONTENTS

I.	INTRODUCTION AND QUALIFICATIONS	3
A.	Qualifications and Professional Experience	3
B.	Scope of this Declaration	5
C.	Person of Ordinary Skill in the Art.....	6
D.	Evidence Considered in Forming My Opinion	7
II.	AUTHENTICITY AND PUBLIC AVAILABILITY OF <i>CURTOTTI</i> (EX. 1008)	7
A.	Authentication	7
B.	Public Availability of <i>Curtotti</i>	9
C.	Usage Records	13
D.	Summary of My Opinion on <i>Curtotti</i>	13
III.	AUTHENTICITY AND PUBLIC AVAILABILITY OF <i>KWAK</i> (EX. 1011)	14
A.	Authentication	14
B.	Discovery of the arXiv copy of <i>Kwak</i>	17
C.	Discovery of <i>Kwak</i> Archived in February 2023	19
D.	Usage Records	22
E.	Summary of My Opinion on <i>Kwak</i>	27
IV.	CONCLUSION	28

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.
U.S. Patent No. 12,353,917

I. INTRODUCTION AND QUALIFICATIONS

I, Ingrid Hsieh-Yee, declare as follows:

1. I am over the age of eighteen and otherwise competent to make this declaration.

2. I have been retained by counsel for Petitioner Wealth, Inc. (“Wealth.com”) as an expert witness in the above-captioned proceeding. I have been asked to provide my expert opinion regarding the authenticity and public availability of certain prior art references in relation to U.S. Patent No. 12,353,917 (“the ’917 Patent”) (**Ex. 1001**). This is my written declaration.

3. I am being compensated for my time in connection with this IPR at my current consulting rate. I am also being reimbursed for any reasonable expenses associated with my work and testimony in this investigation. My compensation is not contingent on the results of my study, the substance of my opinions, or the outcome of this matter.

A. Qualifications and Professional Experience

4. My complete qualifications and professional experience are described in my academic curriculum vitae (**Appendix A**). The following is a brief summary of my relevant qualifications and professional experience.

5. I was a professor in the Department of Library and Information Science at the Catholic University of America for 32 years and was conferred the title of

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.

U.S. Patent No. 12,353,917

Professor Emerita in December 2022. I have experience working in an academic library, a medical library, and a legislative library. I hold a Ph.D. and a Masters in Library and Information Studies from the University of Wisconsin-Madison.

6. I am an expert on library cataloging and classification and have published two editions on this subject, *Organizing Audiovisual and Electronic Resources for Access: A Cataloging Guide* (2000 and 2006). I taught a variety of courses, including Cataloging and Classification, Advanced Cataloging and Classification, Organization of Internet Resources, Organization of Information, Metadata, Digital Content Creation and Management, Internet Searches and Web Design, Information Literacy Instruction, Advanced Information Retrieval and Analysis Strategies, and The Information Professions in Society. I am familiar with metadata schema design and implementation. In my teaching, I covered the design and implementation of metadata in databases, search engines, digital repositories, digital libraries, and digital archives. I also covered how information organization affects the discovery and access to digital resources on the Internet. My research interests cover cataloging and classification, information organization, metadata, information retrieval, information architecture, digital collections, scholarly communication, social media, user interaction with information systems, and others.

7. I am personally familiar with many online catalogs, databases, and search engines. In preparing for this declaration, I used the following authoritative

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.

U.S. Patent No. 12,353,917

information systems to search for records:

- ACL Anthology (<https://aclanthology.org/>)
- arXiv (<https://arxiv.org/>)
- Google Scholar (<https://scholar.google.com/>)
- Internet Archive (<https://archive.org/>)
- SSRN (Social Science Research Network) of Elsevier
(<https://www.ssrn.com/ssrn/>)

B. Scope of this Declaration

8. I have been asked to offer my opinions on the public availability dates of the following documents:

- (1) Curtotti, M., & McCreath, E. (2010, December). Corpus based classification of text in Australian contracts. In *Proceedings of the Australasian Language Technology Association Workshop 2010*, pp. 18-26, **Ex. 1008** (“Curtotti”), and
- (2) Kwak, A., Israelsen, J., Morrison, C., Bambauer, D., & Surdeanu, M. (2022, December). Validity assessment of legal will statements as natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6047-6056, **Ex. 1011** (“Kwak”).

9. I am informed by counsel and understand that a given reference is

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.

U.S. Patent No. 12,353,917

“published” upon a satisfactory showing that such a document has been disseminated or otherwise made publicly available to the extent that persons interested and ordinarily skilled in the subject matter or art could locate it exercising reasonable diligence and obtain the document.

10. I have been informed by counsel that the '917 Patent purports to claim November 29, 2023 as its earliest priority date. I have assessed public accessibility of the documents referenced below as of November 29, 2023.

11. I have been informed by counsel that the subject matter of this proceeding generally relates to machine learning.

C. Person of Ordinary Skill in the Art

12. I have been informed by counsel that a “person of ordinary skill in the art at the time of the invention” (POSITA) is a hypothetical person who is presumed to be familiar with the relevant field and its literature at the time of the inventions. I have been informed by counsel that this hypothetical person is also a person of ordinary creativity, capable of understanding the scientific principles and literature applicable to the pertinent field.

13. I have been informed by counsel that a person of ordinary skill in the art, as of November 29, 2023, would have had an undergraduate degree in computer science or a related field, and three to five years of industry experience with artificial intelligence and machine learning. This level of skill is approximate, and more

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.

U.S. Patent No. 12,353,917

education could have compensated for less experience, and vice versa.

14. It is my opinion that such a person would have been actively engaged in learning about the field, possibly through formal instruction and through exploration of online resources such as ACL Anthology, arXiv.org, Social Science Research Network, Internet Archive, institutional repositories, and digital archives. By at least November 29, 2023, such a person would have had access to a vast array of print and electronic resources, including at least the documents referenced below.

D. Evidence Considered in Forming My Opinion

15. In the preparation of this declaration, I have reviewed **Ex. 1008** and **Ex. 1011** referenced in paragraph 8 above, and any other documents I reference herein. Each of these is a type of material that experts in my field would reasonably rely upon when forming their opinions. I provided a further description of materials I considered in **Appendix B**.

II. AUTHENTICITY AND PUBLIC AVAILABILITY OF CURTOTTI (EX. 1008)

A. Authentication

16. **Ex. 1008** is a true and accurate copy of “Corpus based classification of text in Australian contracts,” (“*Curtotti*”), by Michael Curtotti and Eric McCreath, in *Proceedings of Australasian Language Technology Association Workshop 2010*, pages 18-26, that I received from counsel. This copy is presented as **Ex. 1008** in this declaration.

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.
U.S. Patent No. 12,353,917

17. Page 1 of **Ex. 1008** shows the title, authors, and abstract. The bottom of this page shows the article was published in 2010 in the “Proceedings of Australasian Language Technology Association Workshop,” pages 18 to 26. **Ex. 1008** shows *Curtotti* has a total of nine pages, including five tables, one figure, 20 footnotes, and 19 references.

18. **Ex. 1007** is a true and correct copy of “Natural language processing for extracting conveyance graphs” by Bateman et al., US 2015/0286630 A1 (Oct. 8, 2015), that I received from counsel. Paragraph 0072 references “Corpus Based Classification of Text in Australian Contracts” by Curtotti and McCreath to explain that “training a given language processing model may be based on a combination of machine-learning algorithms and hand-coded rules.” It also shows *Curtotti* is available at <http://www.aclweb.org/anthology/U10-1005>. To determine when this article became publicly accessible on the Internet, I searched the Wayback Machine of the Internet Archive for records, and found the article at this URL was first archived on August 7, 2017. This archived copy from 2017 is presented as **Appendix 1008A** in this declaration.

19. Page 1 of **Appendix 1008A** is a screen capture of a Wayback Machine calendar that shows the file at <http://www.aclweb.org/anthology/U10-1005> was archived four times between “7 Aug 2017” and “18 Apr 2021” and this particular copy was archived on “AUG 07 2017,” available at

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.

U.S. Patent No. 12,353,917

[https://web.archive.org/web/20170807183515/http://www.aclweb.org/anthology/U](https://web.archive.org/web/20170807183515/http://www.aclweb.org/anthology/U10-1005)

[10-1005](https://web.archive.org/web/20170807183515/http://www.aclweb.org/anthology/U10-1005). Pages 2 to 10 of **Appendix 1008A** are the PDF of this archived copy. I have closely compared **Ex. 1008** and **Appendix 1008A** and found them to have the same content for *Curtotti*.

20. **Ex. 1008** is in a condition that raises no suspicion about its authenticity. Specifically, this copy of *Curtotti* does not appear to miss any intermediate pages, and there are no visible alterations to the document. Furthermore, this copy has the same content as the copy archived by Internet Archive on August 7, 2017 (**Appendix 1008A**). Based on my review and analysis, it is my opinion that **Ex. 1008** is authentic.

B. Public Availability of *Curtotti*

21. **Appendix 1008A** shows the copy of *Curtotti* at <http://www.aclweb.org/anthology/U10-1005>, as cited by Bateman et al. (**Ex. 1007**), became publicly available on the Internet no later than August 7, 2017. Since *Curtotti* is available on the ACL Anthology website, I searched Internet Archive for records of this website to determine whether *Curtotti* was archived before 2017. **Appendix 1008B** is a true and correct copy of the web files I encountered during my research. I personally located, identified, and obtained these files. They are the type of materials experts in my field would reasonably rely upon when forming their opinions.

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.
U.S. Patent No. 12,353,917

22. Internet Archive is a digital library that archives scholarly materials, cultural artifacts, and other web resources on the Internet. It provides the public with free access to its collections. I have used the Internet Archive in the past and am familiar with their practice of assigning an extended URL to an archived web resource in the format of `http://www.archive.org/web/[Year in yyyy][Month in mm][Day in dd][Time code in hh:mm:ss]/[Archived URL]` (also known as an extended URL) to show the point of time the archived file was captured.

23. Internet Archive usually does not archive web resources the moment they appear on the Internet, and a web resource must be available first before it can be archived. The archive date of a file captured by the Internet Archive therefore is typically later than the date when the file first became publicly accessible on the Internet. In other words, the archive date means that the archived file was available before, and definitely no later than, the archive date. For example, if a web file was first archived by the Internet Archive on February 1, 2001, that archive date informs us that the web file was available on the Internet for users to discover no later than February 1, 2001. As of that date, users could use search engines such as Google to do keyword searches or title searches to discover the web file. Another option is to search the URL of a file, an institution, or an event on the Wayback Machine, a service of the Internet Archive, to determine if Internet Archive has archived copies of a web file of interest.

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.
U.S. Patent No. 12,353,917

24. Page 1 of **Appendix 1008B** is a screen capture of a Wayback Machine calendar that shows ACL Anthology is “a digital archive of research papers in computational linguistics” and its website at <http://www.aclweb.org/anthology/> was archived 1,180 times. Page 1 shows this particular copy was archived on “AUG 07 2011,” available at <https://web.archive.org/web/20110807051635/http://www.aclweb.org/anthology/>. Page 2 is the PDF of this archived copy that shows links to various events, and the lower right corner shows several links to ALTA. The link for the 2010 ALTA (“10”) leads to the webpage for the “Proceedings of the Australasian Language Technology Association Workshop 2010.”

25. Page 3 of **Appendix 1008B** is a screen capture of a Wayback Machine calendar that shows the webpage for the “Proceedings of the Australasian Language Technology Association Workshop 2010” at <http://aclweb.org/anthology/U/U10/> was archived 47 times and this particular copy was archived on “AUG 03 2011,” available at <https://web.archive.org/web/20110803053509/http://aclweb.org/anthology/U/U10/>. This webpage shows “Corpus Based Classification of Text in Australian Contracts” by Curtotti and McCreath is listed as “U10-1005.” Page 4 is the PDF of this archived copy and the link for “U10-1005” leads to an archived copy of *Curtotti*.

26. Page 5 of **Appendix 1008B** is a screen capture of a Wayback Machine

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.

U.S. Patent No. 12,353,917

calendar that shows *Curtotti* was archived four times, and this particular copy was archived on “JUL 08 2011,” available at <https://web.archive.org/web/20110708111015/http://aclweb.org/anthology/U/U10/U10-1005.pdf>. Pages 6 to 14 are the PDF of this archived copy. I have closely compared this archived copy with **Ex. 1008** and the archived copy of 2017 (**Appendix 1008A**), and found these three copies to have the same content for *Curtotti*.

27. Taken together, the web files in **Appendix 1008B** show that *Curtotti* was publicly available on the Internet no later than July 8, 2011. As of that date, interested users could use search engines, such as Google, to search for *Curtotti*. They could also explore Internet Archive for archived copies of *Curtotti*. Since ACL Anthology is a digital archive of research papers in computational linguistics, users interested in computational linguistics would likely be aware of this digital archive, which means that they could begin their exploration at the website of ACL Anthology at <http://www.aclweb.org/anthology>, then select the link for the 2010 ALTA event to reach the webpage for the Proceedings of the Australasian Language Technology Association Workshop 2010, and discover a link for *Curtotti* (“U10-1005”) that would lead them to the copy of *Curtotti* archived by the Internet Archive on July 8, 2011. The discovery path presented in **Appendix 1008B** shows that in August 2011 interested users could take three simple steps to discover and access

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.

U.S. Patent No. 12,353,917

Curtotti. Further, the ACL Anthology’s home page provides a search bar for users to “Search the Anthology” “via Google,” “via Searchbench @ DFKI,” and “via AAN @ UMich” (see page 1 of **Appendix 1008B**). It is therefore my opinion that *Curtotti* became publicly discoverable and accessible by July 8, 2011.

C. Usage Records

28. Actual usage of a publication is reflected by other publications that make reference to it. I used Google Scholar and Google Patents to review citations to *Curtotti* and selected citations published between 2013 and 2017 to demonstrate usage. **Appendix 1008C** presents seven verified citations and shows the earliest citation was published in March 2013.

D. Summary of My Opinion on *Curtotti*

29. It is my opinion that *Curtotti* (**Ex. 1008**) is authentic because it contains no signs to raise suspicion about its authenticity. Furthermore, this document has the same content as copies of *Curtotti* archived by Internet Archive on July 8, 2011 (**Appendix 1008B**), and August 7, 2017 (**Appendix 1008A**).

30. Web files on the website of ACL Anthology (**Appendix 1008B**) show that in August 2011 interested users could visit this website, which is a digital archive of research papers in computational linguistics, and follow the conference link for the 2010 ALTA to discover *Curtotti*. The ACL Anthology’s home page also provides a search bar for users to “Search the Anthology” “via Google,” “via

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.

U.S. Patent No. 12,353,917

Searchbench @ DFKI,” and “via AAN @ UMich” (see page 1 of **Appendix 1008B**).

Archived records show that *Curtotti* was archived no later than July 8, 2011, meaning it became publicly discoverable and accessible on the Internet no later than that date. Citation records (**Appendix 1008C**) show the earliest citation was published in March 2013.

31. Based on the documents and records I have reviewed, it is my opinion that *Curtotti* is authentic and the article became publicly discoverable and accessible on the Internet no later than July 8, 2011.

III. AUTHENTICITY AND PUBLIC AVAILABILITY OF KWAK (EX. 1011)

A. Authentication

32. **Ex. 1011** is a true and correct copy of “Validity assessment of legal will statements as natural language inference,” (“*Kwak*”), by Kwak et al. in *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6047-6056, that I received from counsel. This copy is presented as **Ex. 1011** in this declaration.

33. Page 1 of **Ex. 1011** (internal page 6047) shows the title, authors, and abstract, and the bottom of this page shows this article is included in the “Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6047-6056” and the conference took place on “December 7-11, 2022.” This page also shows *Kwak* has a “2022” copyright date and the Association for Computational Linguistics holds the copyright. **Ex. 1011** shows *Kwak* has a total of 10 pages,

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.

U.S. Patent No. 12,353,917

including four figures (some color), four tables, nine footnotes, and 19 references.

34. Since *Kwak* is a conference paper, I searched Internet Archive for records. **Appendix 1011A** is a copy of *Kwak* archived by Internet Archive on February 2, 2023, available at <https://web.archive.org/web/20230202142404/https://aclanthology.org/2022.findings-emnlp.447.pdf>. I personally located, identified, and obtained this archived copy. It is the type of material experts in my field would reasonably rely upon when forming their opinions.

35. Page 1 of **Appendix 1011A** is a screen capture of a Wayback Machine that shows *Kwak* was archived twice and this particular copy was archived on “FEB 02 2023,” available at <https://web.archive.org/web/20230202142404/https://aclanthology.org/2022.findings-emnlp.447.pdf>. Pages 2 to 11 are the full article of *Kwak* in PDF. I have closely compared **Ex. 1011** and **Appendix 1011A**, and found them to have the same content for *Kwak*.

36. My research on *Kwak* discovers a copy of this article on arXiv, which is “a curated research-sharing platform open to anyone” that “offers researchers a broad range of service: article submission, compilation, production, retrieval, search and discovery, web distribution for human readers, and API access for machines, together with content curation and preservation” (About arXiv, at

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.

U.S. Patent No. 12,353,917

<https://info.arxiv.org/about/index.html>). The metadata record for *Kwak* at <https://arxiv.org/abs/2210.16989v1> shows Alice Kwak submitted the article on October 30, 2022 (cited as “[arXiv:2210.16989v1](https://arxiv.org/abs/2210.16989v1)”), and the “View PDF” link leads to the full article at <https://arxiv.org/pdf/2210.16989v1>. To determine when this submitted article became publicly available on the Internet, I searched Internet Archive for records. **Appendix 1011B** is a copy of this article archived by Internet Archive.

37. Page 1 of **Appendix 1011B** is a screen capture of a Wayback Machine calendar that shows <https://arxiv.org/pdf/2210.16989v1> was archived twice and this particular copy was archived on “NOV 01 2022,” available at <https://web.archive.org/web/20221101104607/https://arxiv.org/pdf/2210.16989v1.pdf>. Pages 2 to 11 are the full article of *Kwak* in PDF. Since **Appendix 1011B** is a pre-publication submission to arXiv, an open access archive, by the lead author, it does not include the conference information as **Ex. 1011** and **Appendix 1011A**, which are copies of the published article. Specifically, **Appendix 1011B** is different from **Ex. 1011** and **Appendix 1011A** in that 1) it has a banner that identifies the article as [arXiv:2210.16989v1](https://arxiv.org/abs/2210.16989v1) with a “30 Oct 2022” posting date, 2) pages in **Appendix 1011B** are not numbered, 3) **Appendix 1011B** does not show the conference information, and 4) the grant number on page 10 of **Appendix 1011B** is preceded by a # sign. I have closely compared **Appendix 1011B** with **Ex. 1011** and

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.

U.S. Patent No. 12,353,917

Appendix 1011A, and found these three copies to have the same content for *Kwak*.

38. **Ex. 1011** is in a condition that raises no suspicion about its authenticity. Specifically, this copy of *Kwak* does not appear to miss any intermediate pages, and there are no visible alterations to the document. Furthermore, this copy has the same content as the copy archived by Internet Archive on February 2, 2023 (**Appendix 1011A**), and a copy posted by the lead author to arXiv on October 30, 2022 (**Appendix 1011B**). Based on my review and analysis, it is my opinion that **Ex. 1011** is authentic.

B. Discovery of the arXiv copy of *Kwak*

39. To determine the public availability date of *Kwak* submitted to arXiv by the lead author, I searched Internet Archive for records. **Appendix 1011C** is a true and correct copy of the web files I discovered from Internet Archive. They are the type of materials experts in my field would reasonably rely upon when forming their opinions.

40. The first two pages of **Appendix 1011C** are the arXiv metadata record for *Kwak* at <https://arxiv.org/abs/2210.16989v1> that shows the title, authors, abstract, subject, citations for this article (“arXiv:2210.16989 [cs.CL]” and “arXiv:2210.16989v1 [cs.CL] for this version”). The Comments field shows *Kwak* has 10 pages, includes four figures, and is “[t]o be published in the Findings of the Association for Computational linguistics: EMNLP 2022.” The record includes a

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.
U.S. Patent No. 12,353,917
link to the PDF of *Kwak*.

41. Page 3 shows a screen capture of a Wayback Machine calendar that shows the metadata record for *Kwak* (<https://arxiv.org/abs/2210.16989v1>) was archived once on “NOV 01 2022” and available at <https://web.archive.org/web/20221101102929/https://arxiv.org/abs/2210.16989v1>. The “View PDF” link leads to the full article of *Kwak*.

42. Page 5 is a screen capture of a Wayback Machine calendar that shows <https://arxiv.org/pdf/2210.16989v1> was archived twice and this particular copy was archived on “NOV 01 2022” and available at <https://web.archive.org/web/20221101104607/https://arxiv.org/pdf/2210.16989v1>. Pages 6 to 15 are the PDF of this archived copy. I have closely compared this copy of *Kwak* with the arXiv copy (**Appendix 1011B**), and found them to be the same.

43. Taken together, the files in **Appendix 1011C** show that the lead author of *Kwak* submitted the article to arXiv on October 30, 2022 and indicated the article would be published in the Findings of the Association for Computational Linguistics: EMNLP 2022. This pre-publication copy of *Kwak* was archived by Internet Archive on November 1, 2022, making it publicly discoverable and accessible by that date. These dates inform my opinion that in November 2022 interested users could use search engines such as Google to discover the copy of

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.

U.S. Patent No. 12,353,917

Kwak on arXiv. Or, they could visit arXiv to search for *Kwak*. In addition, by November 1, 2022, interested users would be able to discover and access *Kwak* on Internet Archive.

C. Discovery of *Kwak* Archived in February 2023

44. After the EMNLP 2022 conference, *Kwak* was published in the conference proceedings. In early 2023 interested users had several ways to discover the published copy of *Kwak*. **Appendix 1011D**, **Appendix 1011E**, and **Appendix 1011F** present three discovery paths readily available to interested users in early 2023.

45. **Appendix 1011D** is a true and correct copy of records I encountered while researching on *Kwak* in December 2025. Pages 1 and 2 are a screen capture of the metadata record for *Kwak* at <https://aclanthology.org/2022.findings-emnlp.447/> on the ACL Anthology website. The record shows the title, authors, abstract, PDF link (<https://aclanthology.org/2022.findings-emnlp.447.pdf>) and a presentation by the lead author. It also shows *Kwak* is included in the Findings of the Association for Computational Linguistics: EMNLP 2022, and has a publication date of December 2022. Pages 3 and 4 are the PDF of this metadata record.

46. To determine the first public availability date of this published copy of *Kwak*, I searched Internet Archive for records. Page 5 is a screen capture of a Wayback Machine calendar that shows <https://aclanthology.org/2022.findings->

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.

U.S. Patent No. 12,353,917

[emnlp.447.pdf](#) was archived twice by Internet Archive. Page 6 is a screen capture of a Wayback Machine calendar that shows *Kwak* was archived twice and this particular copy was archived on February 2, 2023. Pages 7 to 16 are the full article of *Kwak*, which is the same as **Appendix 1011A**.

47. **Appendix 1011D** shows that in early 2023 interested users could use the URL of *Kwak* (<https://aclanthology.org/2022.findings-emnlp.447.pdf>) to explore the Wayback Machine to obtain a copy of this article.

48. Another path for discovery is to begin with the ACL Anthology website. ACL Anthology is a digital archive for research publications in the field of computational linguistics, so users interested in this area should know about this important source for scholarly publications. **Appendix 1011E** illustrates how interested users could discover *Kwak* in 2023. Interested users could begin their search for *Kwak* at the website of the ACL Anthology. They would then learn that this website was archived 670 times between “24 May 2021” and “1 Jan 2026” (see Page 1 of **Appendix 1011E**). For information about the EMNLP 2022 conference, they would select the website archived on March 14, 2023, available at <https://web.archive.org/web/20230314133352/https://aclanthology.org/>, and follow the “EMNLP 22” link to reach the webpage for the “Conference on Empirical Methods in Natural Language Processing (2022)” archived on March 18, 2023 (see page 2 of **Appendix 1011E**, available at

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.

U.S. Patent No. 12,353,917

<https://web.archive.org/web/20230318065928/https://aclanthology.org/events/emnlp-2022/>).

A screen search for *Kwak* on this webpage would lead to the entry for *Kwak* that includes links for the PDF, bibliographic record, and abstract (see page 3 of **Appendix 1011E**). The PDF link would then lead to the copy of *Kwak* archived on February 2, 2023.

49. **Appendix 1011E** shows that in early 2023 interested users could begin their search for *Kwak* on the ACL Anthology website, select the conference of interest, verify that *Kwak* is included in the conference proceedings, then select the PDF to discover that *Kwak* was archived on February 2, 2023. This four-step discovery path was readily available to interested users in early 2023.

50. Interested users had at least one more path to discover *Kwak* in 2023. They could begin on the ACL Anthology website (see page 1 of **Appendix 1011F** for a screen shot of the website archived on March 14, 2023), and follow the link for “Findings 22” to reach the webpage for the “Findings of the Association for Computational Linguistics (2022)” (see page 2 of **Appendix 1011F** for a screen shot of the webpage archived on March 7, 2023). By doing a screen search for *Kwak*, interested users would discover an entry for *Kwak* with links for the PDF, bibliographic records and abstract (see page 3 of **Appendix 1011F**), then follow the PDF link to reach the copy of *Kwak* archived on February 2, 2023 (see pages 5 to 14 of **Appendix 1011F**).

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.
U.S. Patent No. 12,353,917

51. Another way to discover *Kwak* in 2023 or earlier was to search the ACL Anthology website that stores *Kwak*. A 2018 paper about the ACL Anthology titled “The ACL Anthology: Current State and Future Directions” explains that the ACL Anthology “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)” (see page 1 of **Appendix 1011J**). Further, the ACL Anthology home page archived as of February 2, 2023 showed a search bar (see page 1 of **Appendix 1011F**).

52. Taken together, **Appendix 1011D** shows that interested users could discover the pre-publication copy posted by the lead author on arXiv by November 1, 2022. **Appendix 1011E**, **Appendix 1011F**, and **Appendix 1011J** show that interested users could discover and access the published copy of *Kwak* in early 2023 by visiting the website of ACL Anthology to either (1) select the conference link or conference proceeding link of interest, then do a screen search to find *Kwak* and follow the PDF to access the published copy of *Kwak* archived on February 2, 2023; or (2) search for *Kwak* using the ACL Anthology search functionality (as shown on the home page and as described in **Appendix 1011J**) to locate the published copy of *Kwak*.

D. Usage Records

53. Actual usage of a publication is reflected by other publications that

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.

U.S. Patent No. 12,353,917

make reference to it. I have reviewed citations to *Kwak* on Google Scholar and found one citation to *Kwak* published in 2022: Cao, Y., Sun, Y., Xu, C., Li, C., Du, J., & Lin, H. (2022, December) (**Appendix 1011I**). CAILIE 1.0: A dataset for Challenge of AI in Law-Information Extraction V1. 0. *AI Open*, 3, 208-212. *Kwak* is cited as “Kwak et al., 2022. Kwak Alice Saebom, Israelsen Jacob O., Morrison Clayton T. Validity assessment of legal will statements as natural language inference” (see **Appendix 1011I**, page 14). The publication history of this citing article on page 1 of **Appendix 1011I** shows the manuscript was received by the publisher on June 20, 2022, revised on November 27, 2022, and accepted for publication on December 1, 2022; and the article became available online on December 5, 2022. This 2022 citation supports my opinion that the pre-publication copy of *Kwak* on arXiv (**Appendix 1011B**) became publicly discoverable and accessible no later than November 1, 2022, and the published copy of *Kwak* became publicly discoverable and accessible on ACL Anthology no later than February 2, 2023 (**Appendix 1011A**).

54. Another metric for article usage is the number of downloads an article received over time. The Social Science Research Network (SSRN) provides such data. According to “What is SSRN?” (last updated on August 19, 2021, at <https://www.elsevier.support/ssrn/answer/what-is-ssrn>), “SSRN, formerly known as Social Science Research Network, is an open access research platform used to share

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.

U.S. Patent No. 12,353,917

early-stage research, evolve ideas, measure results, and connect scholars around the world” and this platform “is a searchable online library that enables authors to post their papers and abstracts easily and free of charge. The vast majority of papers can be downloaded free of charge as well.” In addition, “SSRN maximizes usage and discoverability by classifying papers into topic-based eJournals that are distributed to individual and institutional subscribers. This results in downloads and citations that SSRN aggregates into meaningful metrics that authors use for promotion and tenure purposes and that institutions use in their hiring and ranking activities.” Elsevier acquired SSRN in 2016.

55. **Appendix 1011G** is a true and correct copy of the metadata record for *Kwak* on SSRN at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4279728. I personally located, identified, and obtained this record from SSRN. This is the type of record experts in my field would reasonably rely upon when forming their opinions.

56. Pages 1 to 3 of **Appendix 1011G** are screen captures of the SSRN metadata record for *Kwak*. Page 1 shows the title and authors, and indicates *Kwak* is “Arizona Legal Studies Discussion Paper No. 22-21” and is to be published in “Findings of the Association for Computational Linguistics: EMNLP 2022.” It also shows this paper has 11 pages, was first posted on SSRN on November 17, 2022, and last revised on November 21, 2022. Page 2 shows the abstract, keywords, and

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.

U.S. Patent No. 12,353,917

suggested citation for this work, and page 3 shows a link for “Download This Paper,” which leads to the full discussion paper. Page 4 is the cover sheet of this discussion paper that shows Arizona Legal Studies Discussion Paper No. 22-21 was completed in November 2022. Pages 5 to 14 are the full discussion paper. I have closely compared this copy with the pre-publication copy of *Kwak* on arXiv (**Appendix 1011B**), and found them to have the same content, except for the banner of article id and posting date in **Appendix 1011B**. **Appendix 1011G** shows that a pre-publication copy of *Kwak* was posted on SSRN on November 17, 2022, making this copy publicly accessible on SSRN on that date. The “Paper statistics” area on page 1 of **Appendix 1011G** shows this copy of *Kwak* has been downloaded 92 times when I produced this exhibit in late 2025.

57. To have a better understanding of the usage of the copy of *Kwak* on SSRN, I searched Internet Archive for records. **Appendix 1011H** is a true and correct copy of the SSRN metadata record and an author profile archived by Internet Archive. These are the type of records experts in my field would reasonably rely upon when forming their opinions.

58. Page 1 of **Appendix 1011H** is a screen capture of a Wayback Machine calendar that shows the SSRN metadata record for *Kwak* was archived eight times and this particular copy was archived on “NOV 28 2022,” available at <https://web.archive.org/web/20221128060841/https://papers.ssrn.com/sol3/papers.>

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.

U.S. Patent No. 12,353,917

[cfm?abstract_id=4279728](#). The “Paper statistics” area on page 1 shows *Kwak* was downloaded 22 times. Pages 2 and 3 are the PDF of this archived record.

59. Pages 4 to 6 are screen captures of an author profile for Alice Kwak, and the Wayback Machine calendar shows the author profile was archived once by Internet Archive on February 2, 2023. The profile shows Alice Kwak had one scholarly paper on SSRN, which was downloaded 37 times. Page 5 shows the scholarly paper is entitled “Validity Assessment of Legal Will Statements as Natural Language Inference,” which is Arizona Legal Studies Discussion Paper No. 22-21. It also shows the paper is published in Findings of the Association for Computational Linguistics: EMNLP 2022 (forthcoming 2022), and was posted on SSRN on November 17, 2022. The download number is listed as 37 as of February 2, 2023 when this profile was archived by Internet Archive. Pages 7 to 9 are the PDF of this archived author profile.

60. The download numbers presented in **Appendix 1011H** inform my opinion that *Kwak* was downloaded 22 times between November 17, 2022 (when the copy was posted on SSRN) and November 28, 2022 (when the SSRN metadata record for *Kwak* was archived by Internet Archive); and the paper was downloaded 15 more times between November 29, 2022 and February 2, 2023 (when the author profile of Alice Kwak was archived by Internet Archive).

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.
U.S. Patent No. 12,353,917

E. Summary of My Opinion on *Kwak*

61. It is my opinion that *Kwak* (**Ex. 1011**) is authentic because it contains no signs to raise suspicion about its authenticity. Furthermore, this document has the same content as a copy archived by Internet Archive on February 2, 2023 (**Appendix 1011A**) and a copy posted on arXiv on October 30, 2022 (**Appendix 1011B**).

62. *Kwak* has four manifestations—as a published article in a 2022 conference proceedings (**Ex. 1011**), as a copy archived by Internet Archive (**Appendix 1011A**), as a pre-publication copy on arXiv (**Appendix 1011B**), and as a discussion paper on SSRN (**Appendix 1011G**). Archived records show the lead author, Alice Kwak, posted a copy of *Kwak* on arXiv in October 2022, and that copy was archived by Internet Archive on November 1, 2022 (**Appendix 1011C**), making the copy of *Kwak* on arXiv publicly accessible no later than November 1, 2022.

63. Archived records from Internet Archive show the published version of *Kwak* (**Ex. 1011**) was archived by Internet Archive on February 2, 2023 (**Appendix 1011A**), making it publicly discoverable and accessible no later than that date. In early 2023, interested users could discover and access this file by visiting the ACL Anthology website and follow conference links to discover and access *Kwak*, or by searching for *Kwak* using the ACL Anthology's search features (**Appendix 1011E**, **Appendix 1011F**, and **Appendix 1011J**). Citation records show a citing paper to

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.

U.S. Patent No. 12,353,917

Kwak was published in December 2022 (**Appendix 1011I**); the SSRN metadata record for *Kwak* archived in November 2022 shows 22 downloads of the paper; and the SSRN author profile for Alice Kwak archived by Internet Archive in February 2023 shows the paper was downloaded 37 times (**Appendix 1011H**).

64. Based on the documents and records I have reviewed, it is my opinion that the copy of *Kwak* on arXiv became publicly discoverable and accessible no later than November 1, 2022, and a copy of *Kwak* posted on SSRN became publicly discoverable and accessible no later than November 17, 2022. The copy of *Kwak* in a published conference proceedings became publicly discoverable and accessible no later than February 2, 2023 (**Appendix 1011A**). A citing paper to *Kwak* appeared in December 2022.

IV. CONCLUSION

65. In signing this Declaration, I recognize that the Declaration will be filed as evidence in a contested case before the Patent Trial and Appeal Board of the United States Patent and Trademark Office. I also recognize that I may be subject to cross-examination in this proceeding. If required, I will appear for cross-examination at the appropriate time.

66. I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true, and further that these statements were made with the knowledge that willful

Declaration of Dr. Ingrid Hsieh-Yee, Ph.D.

U.S. Patent No. 12,353,917

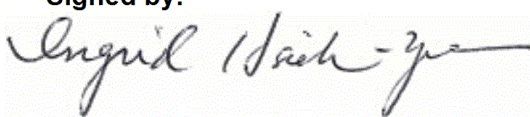
false statements and the like so made are punishable by fine or imprisonment, or

both, under 18 U.S.C. §1001.

Dated: 2/13/2026

Respectfully submitted,

Signed by:



39E37127A44440F...

Ingrid Hsieh-Yee, Ph.D.

Annapolis, Maryland

APPENDIX A

Ingrid Hsieh-Yee

Professor Emerita
Dept. of Library and Information Science
Catholic University of America
Washington, D.C. 20064
E-mail: hsiehyee@cua.edu

Education

Ph.D. 1990. Library and Information Studies, University of Wisconsin-Madison
Minors: Sociology and Psychology

M.A. 1984. Library and Information Studies, University of Wisconsin-Madison.

M.A. 1983. Comparative Literature, University of Wisconsin-Madison.

B.A. 1980. Foreign Languages and Literature, National Taiwan University.

Work Experience

Professor, School/Dept. of Library and Information Science, Catholic University of America,
1990-2022 (Assistant Professor, 1990-1996; Associate Professor, 1997-2004; Professor
2004-2022; Professor Emerita 2022)

Co-Chair, Dept. of Library and Information Science, Catholic University of America, June 2015-
August 2016.

Acting Dean, School of Library and Information Science, Catholic University of America,
January 2010-June 2012.

Cataloger, Dept. of Legislative Reference Library, Annapolis, Maryland, 1989-1990.

Lecturer, School of Library and Information Studies, University of Wisconsin-Madison, 1988.

Teaching Assistant, School of Library and Information Studies, University of Wisconsin-
Madison, 1986-1988.

Cataloger, Health Sciences Library, University of Wisconsin-Madison, 1984-1986.

Areas of Teaching and Research Interests

Information Organization and Access; Metadata; Cataloging & Classification; Information
Architecture; Information Retrieval; Digital Collections; Social Media; Scholarly

Communication; Information Behavior; Health Informatics; Human Computer Interaction; Usability Studies

Grants & Honors

Cultural Heritage Information Management Project. IMLS grant. Amount: \$498,741. Period: Aug. 2012 to July 2015. Co-PI with Dr. Youngok Choi.

D.C. Health Information Technology (HIT4): Building Capacity & Providing Access in Our Nation's Capital. Dept. of Labor H2B Training Grant. Grant amount: \$4,175,500. Grant period: Nov. 2011 to Dec. 2015. Partner with the Metropolitan School of Professional Studies of the Catholic University of America, Children's National Medical Center, D.C. Department of Employment Services, Holy Cross Hospital, Howard University, Center for Urban Progress, Providence Hospital, and Sibley Memorial Hospital.

Capital Health Careers Project. Department of Labor Healthcare Sector and Other High Growth and Emerging Industries Grant. Grant amount: \$4,953,999. Grant period: March 2010 – February 2013. Awarded to a group of healthcare organizations and educational institutions in Washington, D.C. Providence Health Foundation of Providence Hospital (Lead institution). Part of the grant supported the development of a Master's degree program in Information Technology with a concentration in Health Information Technology offered by the School of Library and Information Science.

The Washington D.C. School Librarians Project. IMLS grant. Grant amount: \$412,660. Grant period: Aug. 2007 – June 2011. The School partnered with the District of Columbia Public Schools (DCPS) and the District of Columbia Library Association to educate and mentor school media specialists for the DCPS system. PI, Jan. 2010 to June 2011.

SIG Member of the Year, American Society for Information Science and Technology (2009).

Most Outstanding Paper of *OCLC Systems & Services* (2001).

ALISE Research Grant (2001).

Most Outstanding Paper of *OCLC Systems & Services* (2000).

Research Grant from ERIC (1999-2000).

Best Research Paper Award; Association for Library and Information Science Education (1998).

Research Grants, Catholic University of America. 1991, 1992, 1993, 1996, 1998, 1999, 2004, 2005, 2006, 2007, 2013-14.

Cooperative Faculty Research Grant, Consortium of Universities in the Washington Metropolitan Area (1993-1994).

Cooperative Research Grant, Council on Library Resources (1993-1994).

Journal of the American Society for Information Science Best Paper Award (1993).

ASIS/ISI Information Science Doctoral Dissertation Scholarship (1989).

HEA Title IIB Fellowship (Dept. of Education) (1989)

Chinese-American Librarians Association Scholarship (1987).

Beta Phi Mu (1985).

Vilas Fellowship, University of Wisconsin-Madison. 1984

Publications

Hsieh-Yee, I. (2021). Can We Trust Social Media? *Internet Reference Services Quarterly*, 1-15.
Published online: 28 July, 2021. 10.1080/10875301.2021.1947433.

Bailey, T., & Hsieh-Yee, I. (2020). Combating the Sharing of False Information: History, Framework & Literacy Strategies, *Internet Reference Services Quarterly*, 24 (1-2), 9-30.

Choi, Y., and Hsieh-Yee, I. (2010). Finding Images in an OPAC: Analysis of User Queries, Subject Headings, and Description Notes. *Canadian Journal of Information and Library Science*, 34(3): 271 – 295.

Hsieh-Yee, I. (2008). Educating Cataloging Professionals in a Changing Information Environment. *Journal of Education for Library and Information Science*, 46(2): 93-106.

Vellucci, S. L., Hsieh-Yee, I., and Moen, W.E. (2007). The Metadata Education and Research Information Commons (MERIC): A Collaborative Teaching and Research Initiative. *Education for Information*, 25(3&4): 169-178.

NISO Framework for Guidelines for Building Good Digital Collections. 3rd ed. Baltimore, MD: National Information Standards Organization, 2007. Also available online: <http://www.niso.org/framework/framework3.pdf> (NISO Working Group members: Priscilla Caplan (chair), Grace Agnew, Murtha Baca, Tony Gill, Carl Fleischhauer, Ingrid Hsieh-Yee, Jill Koelling, and Christie Stephenson.)

Choi, Y., Hsieh-Yee, I., and Kules, B. (2007). Retrieval Effectiveness of TOC and LCSH. *Proceedings of the Joint Conference on Digital Libraries*, pp. 233-234.

Vellucci, S., and Hsieh-Yee, I. (2007). They Didn't Teach Me That in Library School! Building

- a Digital Teaching Commons to Enhance Metadata Teaching, Learning and Research. *Proceedings of the National Conference of the Association of College and Research Libraries, Baltimore, MD*, pp. 26-31.
- Mitchell, Vanessa, and Ingrid Hsieh-Yee. (2007). Converting Ulrich's Subject Headings to FAST Headings: A Feasibility Study. *Cataloging & Classification Quarterly*, 45(1): 59-85.
- Hsieh-Yee, I., Tang, R., and Zhang, S. (2007). User Perceptions of a Federated Search System. *IEEE Technical Committee on Digital Libraries Bulletin*, Summer 3(2) (URL = <http://www.ieee-tcdl.org>).
- Tang, R., Hsieh-Yee, I., and Zhang, S. (2007). User Perceptions of MetaLib Combined Search: An Investigation of How Users Make Sense of Federated Searching." *Internet Reference Services Quarterly*, 12(12): 211-236.
- Hsieh-Yee, I., Tang, R., and Zhang, S. (2006). User Perceptions of a Federated Search System. *Proceedings of the Joint Conference on Digital Libraries, June 11-15, 2006, Chapel Hill*, p. 338.
- Hsieh-Yee, I. (2006). *Organizing Audiovisual and Electronic Resources for Access: A Cataloging Guide*. 2nd ed. Westport, Conn.: Libraries Unlimited.
- NISO A Framework of Guidance for Building Good Digital Collections*. 2nd ed. Bethesda, MD: National Information Standards Organization, 2004. Framework Advisory Group: Grace Agnew, Liz Bishoff, Priscilla Caplan (Chair), Rebecca Gunther and Ingrid Hsieh-Yee.
- Hsieh-Yee, I. (2004). Cataloging and Metadata Education in North American LIS Programs. *Library Resources & Technical Services*, 48(1): 59-68.
- Hsieh-Yee, I. (2004). Cataloging and Metadata Education. In Gary E. Gorman (Ed.), *International Yearbook of Library and Information Management 2003: Metadata Applications and Management*, (pp.204-234). London: Facet Publishing.
- Yee, P. L., Hsieh-Yee, I., Pierce, G.R., Grome, R., and Schantz, L. (2004). Self-Evaluative Intrusive Thoughts Impede Successful Searching on the Internet. *Computers in Human Behavior*, 20(1): 85-101.
- Hsieh-Yee, I. (2003). Cataloging and Metadata Education: A Proposal for Preparing Cataloging Professionals of the 21st Century. A report submitted to the ALCTS-Education Task Force in response to Action Item 5.1 of the *Bibliographic control of Web Resources: A Library of Congress Action Plan*. Approved by the Association for Library Collections and Technical Services. Web version available since April 2003 at <http://lcweb.loc.gov/catdir/bibcontrol/CatalogingandMetadataEducation.pdf>.

- Hsieh-Yee, I. (2002). Cataloging and Metadata Education: Asserting a Central Role in Information Organization. *Cataloging & Classification Quarterly* 34(½): 203-222.
- Hsieh-Yee, I., and Smith, M. (2001). The CORC Experience: Survey of Founding Libraries, Part I. *OCLC Systems & Services*, 17: 133-140. (Received "The Most Outstanding Paper of OCLC Systems & Services in 2001" award.)
- Hsieh-Yee, I., and Smith, M. (2001). The CORC Experience: Survey of Founding Libraries, Part II, Automated Tools and Usage. *OCLC Systems & Services*, 17: 166-177. (Received "The Most Outstanding Paper of OCLC Systems & Services in 2001" award.)
- Hsieh-Yee, I. (2001). ERIC User Services: Changes and Evaluation for the Future. *Government Information Quarterly*, 18: 31-42.
- Hsieh-Yee, Ingrid. (2001). Research on Web Search Behavior. *Library and Information Science Research*, 23: 167-185.
- Logan, E., and Hsieh-Yee, I. (2001). Library and Information Science Education in the Nineties. *Annual Review of Information Science and Technology*, 35: 425-477.
- Hsieh-Yee, I. (Ed.) (2001). *Library and Information Science Research*, 23 (2). A special issue in honor of the retirement of Douglas L. Zweigig.
- Hsieh-Yee, I. (2000). *ERIC User Services: Evaluation in a Decentralized Environment*. Washington, D.C.: Dept. of Education.
- Hsieh-Yee, Ingrid. (2000). *Organizing Audiovisual and Electronic Resources for Access: A Cataloging Guide*. Littleton, CO: Libraries Unlimited.
- Hsieh-Yee, I. (2000). Organizing Internet Resources: Teaching Cataloging Standards and Beyond. *OCLC Systems & Services*, 16: 130-143. (Received "The Most Outstanding Paper of OCLC Systems & Services in 2000" award.)
- Hsieh-Yee, I. (1998). The Retrieval Power of Selected Search Engines: How Well Do They Address General Reference Questions and Subject Questions? *Reference Librarian*, 60: 27-47.
- Hsieh-Yee, I. (1998). Search Tactics of Web Users in Searching for Texts, Graphics, Known Items and Subjects: A Search Simulation Study. *Reference Librarian*, 60: 61-85. (Received the 1997 Best ALISE Research Paper Award.)
- Hsieh-Yee, I. (1997). Access to OCLC and Internet Resources: LIS Educators' Views and Teaching Practices. *RQ*, 36: 569-86.

- Hsieh-Yee, I. (1997). Teaching Online and CD-ROM Resources: LIS Educators' Views and Practices. *Journal of Education for Library and Information Science*, 38: 14-34.
- Hsieh-Yee, I. (1996). The Cataloging Practices of Special Libraries and Their Relationship with OCLC. *Special Libraries*, 87: 10-20.
- Hsieh-Yee, I. (1996). Modifying Cataloging Practice and OCLC Infrastructure for Effective Organization of Internet Resources. In *Proceedings of the OCLC Internet Cataloging Colloquium*. [Online]. Available: <http://www.oclc.org/oclc/man/colloq/hsieh.htm>
- Hsieh-Yee, I. (1996). Student Use of Online Catalogs and Other Information Channels. *College & Research Libraries*, 57: 161-175.
- Hsieh-Yee, I. (1995). Ten entries in James S. C. Hu (Ed.), *Encyclopedia of Library & Information Science*, 913, 1028-29, 1036, 1037, 1145-46, 1514, 1575, 1763-64, 2216-27, 2378-79. Taipei, Taiwan: Sino-American Publishing. (Topics include "Advanced Technology/Libraries," "Information Ethics," "Instruction on Cataloging and Classification," "Instruction on Reference Services.")
- Hsieh-Yee, I. (1993). Effects of Search Experience and Subject Knowledge on Online Search Behavior: Measuring the Search Tactics of Novice and Experienced Searchers. *Journal of the American Society for Information Science*, 44: 161-174. (Received the 1993 Best JASIS Paper Award.)

Presentations

- Bailey, T. & Hsieh-Yee, I. (February 2020). The Phenomena of Sharing Misinformation and the Need for Information Literacy. Presented at the 2020 Bridging the Spectrum Symposium, Washington, D.C., February 14, 2020.
- Hsieh-Yee, I. and Fragan-Fly, J. (May 2018) Trends, Design & Strategies for Digital Scholarship Services. Presented at the 2018 Maryland/Delaware Library Association Conference, Cambridge, MD.
- Hsieh-Yee, I. (February, 2018) Research Data Management: What It Takes to Succeed. Presented at the 10th Bridging the Spectrum Symposium, Washington, D.C.
- Hsieh-Yee, I. (February, 2017) *Research Data Management: New Competencies and Opportunities for Information Professionals*. Presented at the 9th Bridging the Spectrum Symposium, Washington, D.C.
- Hsieh-Yee, I. and Lawton, P. (February, 2017) *Enhancing Catholic Portal Searches with User Terms and LCSH*. Presented at the 9th Bridging the Spectrum Symposium, Washington, D.C.

- Hsieh-Yee, I. (2016, October) *Visualizing Data for Information*. Presented at the 2016 Virginia Library Association Conference, Hot Springs, VA.
- Hsieh-Yee, I. (2016, August) *Religious Materials Toolbox for Archivists: Solutions to Problems Facing the Profession*. Presented at Archives * Records 2016, Atlanta, GA.
- Hsieh-Yee, I. and Lawton, P. (2016, March) *Enhancing Retrieval of Catholic Materials with LCSH Knowledge Structure*. Presented at the 2016 Catholic Library Association Conference, San Diego, CA.
- Fagan-Fry, J. and Hsieh-Yee, I. (2016, February) *Approaches to Digital Scholarship at Top Universities around the World: Scholarly Publishing in the Digital Age*. Presented at the 8th Bridging the Spectrum Symposium, Washington, D.C.
- Hsieh-Yee and Fagan-Fry, J. (2016, January) *Innovative Services for Digital Scholarship at Top 100 Research Libraries of the World*. Poster presented at the 2016 Annual Conference of the Association for Library and Information Science Education, Boston, Mass.
- Hsieh-Yee, I. and Lawton, P. (2015, June). *Crowdsourcing terms for CRRA portal themes*. Poster presented at the third CRRA symposium and annual meeting, Bringing the created toward the Creator: Liturgical art and design since Vatican II. Catholic Theological Union, Chicago, Illinois.
- Hsieh-Yee, I. and Lawton, P. (2015, February). *Crowdsourcing terms for thematic exploration in the Catholic Portal*. Poster presented at the 7th Annual Bridging the Spectrum Symposium, Washington, D.C.
- Hsieh-Yee, I., James, R., and Fagan-Fry, J. (2015, February). *Support for digital scholarship at top university libraries of the world*. Poster presented at the 7th Annual Bridging the Spectrum Symposium, Washington, D.C.
- Hsieh-Yee, I., Zhang, S., Lin, K., and Cherry, S. (2015, February). *Thus said the end users: Summon experience and support for research workflows*. Poster presented at the 7th Annual Bridging the Spectrum Symposium, Washington, D.C.
- Yontz, E., Hsieh-Yee, I., & Houston, S. (2015, February). *Healthy Heroes Summer Reading Club: Developing healthy youth at public libraries*. 11th Annual Jean Mills Health Symposium, Greenville, North Carolina.
- Yontz, E., Hsieh-Yee, I., and Houston, S. (2015, January). *Healthy youth and libraries: A pilot study*. Association for Library & Information Science Education (ALISE) Annual Conference, Chicago, Illinois.

- Hsieh-Yee, I. (2014, May). *Linking CRRA resources to portal themes via authority files*. Presented at the Catholic Research Resources Alliance 2014 Membership Meeting, Marquette, WI.
- Hsieh-Yee, I. (2014, April). *Enhancing subject access to CRRA resources*. Presented at the 2014 Catholic Library Association Conference, Pittsburgh, PA.
- Hsieh-Yee, I. (2014, January). *Health Information Technology Program: Educational entrepreneurship in action*. Presented at the 2014 annual Conference of the Association for Library and Information Science Education, Philadelphia, PA
- Hsieh-Yee, I., Zhang, S., Lin, K., and Cherry, S. (2014, January). *Discovering information through Summon: An analysis of user search strategies and search success*. Paper presented at the 6th Bridging the Spectrum Symposium, Washington, D.C.
- Hsieh-Yee, I. (2012, December). *National Digital Stewardship Alliance and SLIS at CUA: An Educational Partnership*. Paper presented at Best Practices Exchange: Acquiring, Preserving, and Providing Access to Government Information in the Digital Era, Annapolis, MD
- Choi, Y. and Hsieh-Yee, I. (2010, January). *Finding Images in an OPAC: Analysis of User Queries, Subject Headings, and Description Notes*. Paper presented at 2nd Annual Bridging the Spectrum Symposium, Catholic University of America, Washington, D.C.
- Hsieh-Yee, I. and Coogan, J. (2010, January). *Google Scholar vs. Academic Search Premier: What Libraries and Searchers Need to Know*. Paper presented at 2nd Annual Bridging the Spectrum Symposium, Catholic University of America, Washington, D.C.
- Hsieh-Yee, I. (2009, November). *Information Science Education: An LIS School's Perspective*. Paper presented at Annual Meeting of the American Society for Information Science and Technology, Vancouver, British Columbia, Canada.
- Hsieh-Yee, I., Menard, E., Ya-Ning Chen, A., Shu-Jiun Chen, S., Kalfatovic, M. R., Wisser, K. M. (2009, November). *Information Organization in Libraries, Archives and Museums: Converging Practices and Collaboration Opportunities*. Presented at Annual Meeting of the American Society for Information Science and Technology, Vancouver, British Columbia, Canada. (Organizer and moderator of this panel.)
- Hsieh-Yee, I. and Coogan, J. (2009, July). *Catching up to Google Scholar: The Retrieval Power of Academic Search Premier and Google Scholar*. Poster presented at American Library Association Conference, Chicago, Illinois.
- Hsieh-Yee, I., with the CUA Scholarly Communications Project Team. (2009, January). *Digital Scholarship@CUA: Developing an Institutional Repository for CUA*. Poster presented at

1st Annual Bridging the Spectrum Symposium, Catholic University of America, Washington, D.C.

Wise, M., Cylke, K., and Hsieh-Yee, I. (2009, January). *Digital Talking Books: Meeting the Needs of the Blind and the Handicapped*. Paper presented at the Bridging the Spectrum Symposium, Catholic University of America, Washington, D.C.

Hsieh-Yee, I. (2009, January). *User Expectations of MERIC*. Presented at the Information Organization Competencies for the 21st Century Discussion Session of the 2009 Conference of the Association for Library and Information Science Education, Denver, Colorado.

Choi, Y., and Hsieh-Yee, I. (2008, November). *Subject Access for Images in an OPAC*. Annual Meeting of the American Society for Information Science and Technology, Columbus, Ohio. (Also co-organized a panel on Retrieving and Using Visual Resources: Challenges and Opportunities for Research and Education.)

Hsieh-Yee, I. (2008, June). *Educating Cataloging Professionals in a Changing Information Environment*. National Taiwan University, Taipei, Taiwan.

Vellucci, S. L., Moen, W.E., Hsieh-Yee, I., Marson, B., and Wisser, K. (2008, January) *Building a Metadata Education and Research Community through MERIC (Metadata Education and Research Information Commons): Demo and Stakeholder Input*. A panel presented at the 2008 Conference of the Association for Library and Information Science Education, Philadelphia, Pennsylvania.

Hsieh-Yee, I., Choi, Y. and Kules, B. (2007, October). *Searching for Books and Images in OPAC: Effects of LCSH, TOC and Subject Domains*. A poster presented at the American Society for Information Science and Technology Annual Meeting, Milwaukee, Wisconsin.

Hsieh-Yee, I. and Coogan, J. (2007, August) *Google Scholar vs. Academic Search Premier: A Comparative Analysis*. Presented to the Faculty and Staff of the University of the District of Columbia.

Hsieh-Yee, I. and Coogan, J. (2007, June). *Google Scholar vs. Academic Search Premier: A Comparative Analysis*. Presented to the Washington Research Library Consortium Community, Catholic University of America, Washington, D.C.

Hsieh-Yee, I., Choi, Y., and Kules, B.. (2007, June). *What Users Need for Subject Access: Table of Contents or Subject Headings?* A poster presented at the 2007 American Library Association Annual Conference, Washington, D.C., June 2007.

Choi, Y., Hsieh-Yee, I., and Kules, B. (2007, June). *Retrieval Effectiveness of TOC and LCSH*. A paper presented at the Joint Conference on Digital Libraries 2007, Vancouver, Canada.

- Vellucci, S. L., Hsieh-Yee, I., and Moen, W.E. (2007, May). *If We Build It, Will They Come? Building a Community of Practice for Metadata Stakeholders*. A poster presented at the Rutgers University Research Day, Bridgeton, New Jersey.
- Hsieh-Yee, I. (2007, May). *Federated Searching: User Experience & Perceptions*. International Conference on Information Organization & Retrieval, National Taiwan University, Taipei, Taiwan.
- Hsieh-Yee, I. (2007, May). *Search Performance of Google Scholar and Academic Search Premier*. International Conference on Information Organization & Retrieval, National Taiwan University, Taipei, Taiwan.
- Hsieh-Yee, I. (2007, May) *MERIC: Building a Digital Commons for Metadata Education & Research*. International Conference on Information Organization & Retrieval, National Taiwan University, Taipei, Taiwan.
- Hsieh-Yee, I., and Coogan, J. (2007, March/April). *A Comparative Analysis of Google Scholar and Academic Search Premier*. Poster presented at the Association of College & Research Libraries 13th National Conference, Baltimore, Maryland.
- Vellucci, S. L. and Hsieh-Yee, I. (2007, March/April) *They Didn't Teach Me That in Library School! Building a Digital Teaching Commons to Enhance Metadata Teaching, Learning and Research*. On-site presentation and Webcast by Elluminate. A contributed paper presented at the Association of College & Research Libraries 13th National Conference, Baltimore, Maryland. The acceptance rate for contributed paper was 20%. This paper was one of 10 conference papers chosen for live webcast during the conference.
- Moen, W., Hsieh-Yee, I. and Vellucci, S.L. (2007, January) *A DSpace Foundation for a Teaching & Research Commons: The Metadata Education and Research Information Commons*. A poster session presented at the Open Repositories Conference 2007, San Antonio, Texas.
- Tang, R., Hsieh-Yee, I., and Zhang, S. (2006, November) *User Perception of MetaLib Combined Search*. Paper presented at the Annual Meeting of the American Society for Information Science and Technology, Austin, Texas, Nov. 2006.
- Hsieh-Yee, I. (2006, November). *Federated Searching: User Perceptions, System Design, and Library Instructions*. Paper presented at the Annual Meeting of the American Society for Information Science and Technology, Austin, Texas. (Panel organizer, moderator, presenter).
- Hsieh-Yee, I. (2006, November). *Building a Digital Teaching Commons to Enhance Teaching and Learning: The MERIC Experience and Challenges*. Paper presented at the Annual Meeting of the American Society for Information Science and Technology, Austin,

Texas. (Panel organizer, moderator, presenter)

Hsieh-Yee, I. (2006, September). *Search Performance of Google Scholar and Academic Search Premier*. Paper presented at the ERIC Publishers Meeting, Washington, D.C.

Hsieh-Yee, I., Zhang, S., and Rong Tang, R. (2006, June). *User Perceptions of a Federated Search System*. Poster presented at Joint Conference on Digital Libraries, Chapel Hill, North Carolina.

Hsieh-Yee, I. and Zhang, S. (2006, June). *Preparing Users for Federated Search: Implications of a MetaLib User Perceptions Study*. Paper presented at the 2006 Ex Libris User Groups of North America Conference, Knoxville, Tennessee.

Hsieh-Yee, I. (2006, January). *MERIC Organizations and Navigation*. Paper presented at the 2006 ALISE Annual Conference, San Antonio, Texas.

Hsieh-Yee, I. (2006, January). *Metadata and Cataloging Education: Recommended Competencies*. Paper presented at the 2006 ALISE Annual Conference, San Antonio, Texas.

Hsieh-Yee, I. (2005, November). *Digital Library Evaluation: Progress & Next Steps*. Presentation at the Annual Meeting of the American Society for Information Science & Technology, Charlotte, North Carolina.

Hsieh-Yee, I. (2005, August). *Providing Access to Digital Content: Issues for DL Managers*. Presentation at MDK12 Digital Library Steering Committee Meeting, Columbia, Maryland.

Hsieh-Yee, I. (2005, April). *Enhancing Teaching and Learning: The Role of School Library Media Specialists*. Presentation at Meeting of the Baltimore County Public School System School Media Specialists, Baltimore, Maryland.

Hsieh-Yee, I. (2005, January). *Subject Access and Users: Insights & Inspirations from Marcia J. Bates*. Paper presented at the Historical Perspectives SIG, 2005 Conference of the Association for Library and Information Science Education, Boston, Massachusetts.

Hsieh-Yee, I. (2005, January). *Electronic Resource Management: Practice, Employer Expectations, & CE Interests*. Paper presented at Technical Services Education SIG, 2005 Conference of the Association for Library and Information Science Education, Boston, Massachusetts.

Hsieh-Yee, I. (2004, October). *Library Professionals for the Digital Age: Competencies & Preparation*. Paper presented at Bibliographic Access Management Team meeting, Library of Congress, Washington, D.C.

- Hsieh-Yee, I. (2004, January). *Cataloging and metadata expertise for the digital era*. Presented at Preparing 21st Century Cataloging and Metadata Professionals: A Workshop for Educators and Trainers, San Diego and sponsored by ALCTS, ALISE, LC, and OCLC.
- Hsieh-Yee, I. (2004, January). *Educating catalogers for the digital era*. Paper presented at the Technical Services SIG, 2004 Conference of the Association for Library and Information Science Education, San Diego.
- Hsieh-Yee, I. (2003, July). *Cataloging Education for the 21st Century*. A presentation at the Library of Congress, Washington, D.C.
- Hsieh-Yee, I. (2002, January) *Metadata Education and Research Priorities: A Delphi Study of Metadata Experts*. Presentation at the 2002 Conference of the Association for Library and Information Science Education, New Orleans.
- Hsieh-Yee, I. (2001, November). *A Delphi Study of Metadata: Preliminary Findings*. Poster session at the 2001 Annual Meeting of the American Society for Information Science & Technology, Washington, D.C.
- Hsieh-Yee, I. (2001, June). *Resources on Asian American Children: Analysis of Retrieval by Search Engines and WorldCat*. Presentation at the National Conference on Asian Pacific American Librarians, San Francisco.
- Hsieh-Yee, I. (2001, January). *Delphi Study on Metadata: Project Design*. Presentation at Research Awards Session, Association for Library & Information Science Education, Washington, D.C.
- Hsieh-Yee, I. (2000, May). *Web Search Behavior Research: Progress and Implications*. Presentation at the Symposium on Evaluating Library and Information Science Research, University of Wisconsin-Madison, Madison, Wisconsin.
- Hsieh-Yee, I. (2000, March). *ERIC User Services: Evaluation in a Decentralized Environment*. Presentation at the National ERIC Joint Directors/Technical Meeting, Arlington, Virginia.
- Hsieh-Yee, I. (2000, January). *Enhancing Learning with Web Technology*. Presentation at Faculty Conversations, Catholic University of America, Washington, D.C.
- Hsieh-Yee, I. (2000, January). *From Surrogates to Objects: CUA's Approaches to Organizing Electronic Resources*. Paper presentation at the Annual Conference of the Association for Library and Information Science Education, San Antonio, Texas.
- Yee, P., and Hsieh-Yee, I. (1997, November). *Individual Differences in Search Behavior on the WWW*. A poster session presented at the 38th Annual Meeting of the Psychonomic

Society, Philadelphia, Pennsylvania.

Hsieh-Yee, I. (1997, April). *Research + Marketing + Preparation = Job!* Presented at the "Workshop on Resume and Interview Techniques," Special Libraries Association, Student Chapter, Catholic University of America, Washington, D.C.

Hsieh-Yee, I. (1997, February). *Creating CyberCatalogers: Education and Training.* Presentation at ALA's Midwinter Meeting, Washington, D.C.

Hsieh-Yee, I. (1997, February). *Search Tactics of Web Users in Searching for Texts, Graphics, Known Items and Subjects: A Search Simulation Study.* Presented at the Conference of the Association for Library and Information Science Education, Washington, D.C.

Hsieh-Yee, Ingrid. "Beginning Your Special Library/Information Center Career." Presented at SLA's "Career Day," Jan. 11, 1997, Catholic University of America.

Hsieh-Yee, I. (1996, September). *The Roles of Library and Information Scientists in Managing Electronic Information.* Presentation at Hamilton College, Clinton, New York.

Hsieh-Yee, I. (1996, May). *The Future of Cataloging as a Profession.* Presented at "The Cataloging Forum, Library of Congress, Washington, D.C.

Hsieh-Yee, I. (1994, October). *The Impact of the Internet on OPACs.* Presented at the Third Workshop on User Interfaces for OPACs, Library of Congress, Washington, D.C.

Reports

Hsieh-Yee, I., with Knowledge Management Competencies and Performance Action Group of the Federal Knowledge Management Initiative. "From Knowledge Management Competencies to Improved Organizational Performance." April 9, 2009.

Hsieh-Yee, I., with Knowledge Practices Action Group of the Federal Knowledge Management Initiative. "KM Practice in Government Agencies: Findings and Recommendations." April 9, 2009.

Hsieh-Yee, I. "Delphi Study on Metadata." 2001. Three quarterly reports submitted to the Association for Library and Information Science Education.

Hsieh-Yee, I. "College Students' Information Channels: Patterns of Use and Possible Factors in Channel Selection." 1995. Submitted to the Catholic University of America.

Hsieh-Yee, I. "The Information-Seeking Patterns of Scholars and Their Use of an Online Information System." 1994. Submitted to the Council on Library Resources.

Book Reviews

- Review of *The Measurement and Evaluation of Library Services*, by Sharon L. Baker and F. Wilfrid Lancaster. *Information Processing and Management* 30 (1994): 450-52.
- Review of *Subject Access to Films and Videos*, by Sheila S. Intner and William E. Studwell; and *Cataloging Unpublished Nonprint Materials*, by Verna Urbanski with Bao Chu Chang and Bernard L. Karon. *Information Processing and Management* 30 (1994): 449-50.
- Review of *Automated Information Retrieval in Libraries: A Management Handbook*, by Vicki Anders. *Journal of Library and Information Science* 19 (1993): 98-100.
- Review of *Full Text Databases*, by Carol Tenopir and Jung Soon Ro. *Information Processing and Management* 28 (1992): 667-68.
- Review of *Descriptive Cataloging for the AACR2R And USMARC: A How-to-Do It Workbook*, by Larry Millsap and Terry Ellen Ferl. *Information Processing and Management* 28 (1992): 809-11.
- Review of *MARC Manual: Understanding and Using MARC Records*, by Deborah J. Byrne. *Information Processing and Management* 28 (1992): 537-38.

Service

Professional Associations and Societies

- Library of Congress. RDA Training Program for the Profession. Co-authored with Tim Carlton. 2013-2014.
- 2014 Digital Preservation Outreach & Education Survey. Contributed to the design of the survey, 2014.
- National Digital Stewardship Alliance. Outreach Committee. 2011-2014.
- National Digital Stewardship Residency Program. Advisory Group, 2012-2013.
- FEDLINK Health Information Technology Advisory Council, 2011-2015.
- 2012 Joint Conference on Digital Libraries. Program Planning Committee, Pre-Conference Proposals Review Committee, 2012
- Catholic Research Resources Alliance. Five-Year Strategic Plan Task Force, 2011-2012
- Institute of Museum and Library Services. Grant reviewer. 2004, 2005, 2010.
- Association for Library and Information Science Education.
 - * ALISE Bodan Wynar Research Paper Award Committee, 2015, 2016, 2017
 - * ALISE Eugene Garfield Dissertation Award Competition, Jury, 2013, 2014
 - * ALISE Research Grant Competition Committee. Chair, 2012
 - * Pratt-Severn Faculty Innovation Award. Chair, 2009, 2010
 - * ALISE Doctoral Poster Jury, 2012
 - * "Information Organization Competencies for the 21st Century" Discussion session leader. 2009 Conference of the Association for Library and Information Science

Education.

- * Assisted Technical Services SIG Convener in organizing a program, “Building a Metadata Education and Research Community through MERIC (Metadata Education and Research Information Commons): Demo and Stakeholder Input” for the 2008 ALISE conference.

- * Association for Library Collections and Technical Services/Association for Library and Information Science Education (ALCTS/ALISE) Metadata Education and Research Information Center (MERIC) Advisory Board, Co-Chair (with Sherry Vellucci), 2005-2007. Chair, 2008-2009 (leading the effort to build MERIC, a repository and collaborative space for metadata educators, practitioners, and researchers)

- * Technical Services SIG, Convener, 2004-2005. Organized a program on “Electronic Resources Management: Current Practices, Employer Expectations, and Teaching Strategies” for the 2005 conference in Boston, Massachusetts.

- * Technical Services SIG, Convener, 2003-2004. Organized a program on “Organizing Information with Metadata: Desired Competencies and Teaching Innovations” for the 2004 conference.

- * Technical Services SIG, Convener, 1999-2000. Organized a program on "Teaching the Organization of Electronic Resources" for the 2000 conference.

- * Curriculum SIG, Co-convener (with Sibyl Moses), 1996-97. Organized a program on “Government Information Policy” for the 1997 conference.

- American Society for Information Science & Technology.

- * Reviewer, Conference program panel submissions and poster submissions, 2005, 2006, 2007, 2009, 2011, 2012, 2013, 2014, 2015, 2016, 2017

- * Nomination Committee, 2009-2011

- * Information Science Education Special Interest Group. American Society for Information Science and Technology. Chair-Elect, 2007-2008. Chair 2008-2009.

- * Committee on Information Science Education. 1999-2006.

- * Committee on Information Science Education. Organizing Committee for an orientation program for students at ASIS annual meetings, 1999-2001

- * Committee on Information Science Education. Sub-committee on Student Welfare (focusing on issues related to master's education), 1998-2001

- * SIG ED. Organizing Committee for the "Seminar on Research and Career Development" for junior researchers. 1995-96 (chair), 1997-2001

- * ISI Doctoral Dissertation Proposal Scholarship Jury, 1997; 2001, 2002

- * Pratt-Severn Best Student Research Paper Award Jury. Chair. 1997

- * 1998 Midyear Meeting (referee of contributed papers), 1997

- * Organizer and moderator of the ASIS Doctoral Forum and the Doctoral Research Seminar 1994-1995

- * SIG Human Computer Interaction. Chair-Elect, Chair, 1993-1995

- * Doctoral Forum Award Jury, 1995

- * Best Student Paper Award Jury, 1995

- American Library Association.

- * Committee on Accreditation, External Review Panelist, 2009- (site visiting team 2013-2014; site visiting team 2016-2017)
 - * Association for Library Collections and Technical Services Task Force on Competencies and Education for a Career in Cataloging, member, 2008-2009
 - * Facilitator for “What They Don't Teach in Library School: Competencies, Education and Employer Expectations for a Career in Cataloging,” an Association for Library Collections and Technical Services Preconference, June 22, 2007 in Washington, D.C. Also a local liaison for bringing this program to the Catholic University of America.
 - * Facilitator for a discussion on "Effect of Electronic Resources on Technical Services" at ALA's Midwinter Meeting held in Feb. 1997 in Washington, D.C.
 - * International Relations Committee, Subcommittee Task Force for IFLA and China, 1994-1997
- Virginia Association of School Librarians. Scholarships and Awards Committee. 2010-2012
- Federal Knowledge Management Initiative, Knowledge Management Practices Action Group. Member. 2009 (leading the effort to build a knowledge management repository)
 - Federal Knowledge Management Initiative, Knowledge Management Competencies & Learning Action Group. Member. 2009 (developing an action plan for helping government knowledge workers and government agencies to develop knowledge management competencies)
 - National Center for Education Statistics. Technical Review Panel. 2008.
 - External evaluator for a case of promotion to full professorship. University of Tennessee. 2008.
 - National Information Standards Organization (NISO). Advisory Board, Revision of “IMLS Framework of Guidance for Building Good Digital Collections,” 2004, 2007.
 - Library of Congress, Bibliographic Control of Web Resources: A Library of Congress Action Plan. Principal Investigator of Action Item 5.1, focusing on cataloging and metadata education for students and new librarians, 2002-2003. (worked with the Association for Library Collections and Technical Services, Education Task Force)
- Chinese American Librarians Association
 - * Chinese American Librarians Association Outstanding Library Leadership Award in Memory of Dr. Margaret Chang Fung, Award Committee, 2016-2017
 - * Achievement Award Jury, 2000-2001
 - * CALA Goal 2000 Task Force, 1997
 - * Scholarship Committee, 1995, 1996-1997 (chair)
 - * Board of Directors, 1994-1997
 - * Publication Committee, 1993-1995
 - * International Relations Committee, 1993-1996
 - SailorSM Assessment Advisory Group (An impact study of Sailor, Maryland's Public Information Network), 1995

- Editorial boards
 - Journal of Library & Information Science. Editorial Board, 2012-
 - Chinese American Librarians Association, *Occasional Papers Series*. Editorial Board, 2009-2016.
 - Library Quarterly*. Editorial Board, 2003-2008
 - Bulletin of the Medical Library Association*, 1994-97
 - Newsletter editor for the Chinese American Librarians Association, 1989-92
- Referee for the following journals
 - Information Processing and Management*
 - Journal of Digital Information*
 - Journal of Education for Library and Information Science*
 - Journal of Information Science*
 - Journal of Library & Information Science*
 - Journal of Library Metadata*
 - Journal of the American Society for Information Science & Technology*
 - Library and Information Science Research*
 - Library Quarterly*
- Expert reviewer, “Digital Library” course, Evaluation module, University of North Carolina, Chapel Hill, 2007-2008.
- Expert reviewer, “Information Organization” course, University of Michigan, Ann Arbor. 2007.

Catholic University of America

- Academic Senate Committee on Appointments and Promotions, 2005-2008
- Academic Senate Committee on Committees and Rules, 2009-2012, 2019-2022
- Academic Senate representative (for School of Arts & Sciences), 2017-2020
- Academic Senate representative (for School of Library & Information Science) 2003-2012
- Academic Senate Library Committee, Interim Chair (2007), Member, 2008-2012
- Academic Senate Committee on Computing, 1995-2003
- President’s Administrative Council, 2010-2012
- Deans’ Council, 2010-2012
- Academic Leadership Group, 2010-2012
- Graduate Board, 2010-2012
- School of Arts & Sciences, Committee on Appointments and Promotions, 2015-2019
- School of Arts & Sciences, Academic Council, 2015-2016.

- School of Arts & Sciences, Ordinary Professor Group, 2013-
- Doctoral Dissertation Defense Committee, Chair, Dept. of Psychology, 2016, 2017, 2018, 2019, 2020, 2021, 2022
- Doctoral Dissertation Defense Committee, Chair, Dept. of Education, 2014, 2015, 2017, 2018, 2019
- Doctoral Dissertation Defense Committee, Chair, National Catholic School of Social Services, 2019, 2020, 2021
- Doctoral Dissertation Defense Committee, Chair, School of Nursing, 2006, 2008
- CUA Scholarly Communication Project Team, Member (2007), Chair, 2008-2009
- Dean Search Committee, 1992-1994, 1998-1999, 2002-2003, 2006-2007
- Fulbright Review Panel, 2006
- CUA Service Learning Advisory Board, 2001-2002
- CUA Faculty Conversations on Enhancing Teaching and Learning through Technology, Planning Group, 1999-2001
- CUA Initiative on Technology and Teaching, 1998-2001

Dept. of Library and Information Science

- Admissions Committee, 2007-2009, Chair 2010-2012, Member 2013-2015, Member 2018-2022
- Appointments and Promotions Committee, 1991-2022
- Student advisement, 1990-2022
- 4+1 Bachelor to Master program, Program Advisor, 2014-2022
- LIS Minor, Advisor, 2017-2022
- Curriculum Committee, 2020-21 (Member), 2021-2022 (Chair)
- Scholarship & Awards Committee, 2016-2018, 2020-21
- Comprehensive examination editor, 2020-21
- Symposium, Colloquium, Lecture Series Committee, May 2018-May 2020 (Chair)
- Symposium and Colloquium Committee, fall 2016-May 2018
- Bridging the Spectrum Symposium session moderator, 2018, 2019, 2020, 2021
- Organization of Information Core Course Outcomes Assessment, 2019, 2020
- The Information Professions in Society Core Course Outcomes Assessment, 2019, 2020
- Community Services Librarianship, Course of Study Review and Revision, 2020
- Organization of Information, Course of Study Review and Revision, 2019
- Organization of Information, Competency and Course Offerings Review, 2018-2019
- Accreditation presentation, Chair, June 2015-August 2016
- Interim Co-Chair, June 2015-August 2016.
- Blended/OWL Learning Committee, spring 2016-2018
- Technology Committee, fall 2016-2017
- Comprehensive examination editor, 2016-2017, reader (every year since 1990)
- LIS Advisory Board, 2015-2016 (chair); fall 2016- May 2018 (member)
- Committee on Planning and Assessment, 2015-2016 (chair)

- Senior Faculty Committee, 2014-2016.
- Accreditation Steering Committee, 2014-2016 (Chair, 2015-2016)
- Accreditation Students Standard Committee, co-chair, 2014-2016
- Accreditation Mission, Goals, and Objectives Standard Committee, co-chair, 2014-2016
- Accreditation Curriculum Standard, member 2014-2-16
- Accreditation Administration and Finance Standard, member 2014-2016
- Cultural Heritage Information Management Project (IMLS-funded), Co-PI, 2012-2015
- Cultural Heritage Information Management Forum (scheduled for June 2015), Co-Organizer, 2013-2015
- Health Information Technology Interim Review Committee, 2015 (chair)
- Health Sciences Librarianship Advisory Group, 2015- (chair)
- Comprehensive examination editors, 2013-2014, 2016-2017
- National Digital Stewardship Alliance liaison, 2011-2014
- Advisory Board, Chair 2010-2012
- Academic Honesty Committee, Chair, 2008-2012
- Blended Learning Committee, 2010-2012
- Colloquium Committee, 2010-2012
- Comprehensive Examination Administration, 2010-2012
- Cultural Heritage Information Management Advisory Committee, 2010-2012 (chair), 2013-
- Curriculum Committee, 1991-2003, 2007-2009, Chair 2010-2012, member 2013-
- Curriculum Subcommittee on Comprehensive Examination, Chair 2009-2012
- Health Information Technology Advisory Board, Chair 2010-2012. Member 2013-
- Health Sciences Advisory Committee, 2009, Chair 2010-2012. Member 2013-
- HIT Expert Forum, Chair 2012. Member 2013-2015
- Health Information Technology Student Group Advisor, 2011-2012
- State Council for Higher Education of Virginia, SLIS Representative, 2010-2012
- Symposium Planning Committee, 2010-2012
- Website Management Team, Chair, 2010-2012
- Urban School Librarianship Project (IMLS-Funded), PI, 2007-2011 (chair, 2010-11)
- Failing Grades Committee, 1995-1997 (chair), 2000-2001 (chair), 2004-2005 (chair), 2007 (chair)-2011
- Faculty Search Committee, 1994-1998, 2002-2004, 2006 (chair), Fall 2007-2009, Chair fall 2009-2012
- Recruitment Committee, Chair 2010-2012
- Strategic Planning Committee, Chair 2010-2012
- Technology Committee, 2010-2012
- Accreditation Advisory Committee, 2007-2009
- Accreditation Coordinating Committee, 2007-2009
- Accreditation Steering Committee, 2007-2009
- SLIS Advisory Group, 2007-2009
- Accreditation Curriculum Standard Committee, Co-chair, 2007-2009
- Accreditation Faculty Standard Committee, Co-chair, 2007-2009

- LSC 551 Information Organization Review Team, Co-chair, 2008-2009, 2015-2016.
- Curriculum Subcommittee on Portfolios, 2009
- LSC 555 Information Systems in Libraries and Information Centers Review Team, contributor, 2008-2009
- Redesign of LSC 730 Use and Users of Libraries and Information. 2009-
- Development of a metadata institute that was taught as LSC 715 Organization of Internet Resources in 2008. The institute is being revised and will be offered in 2010 under a new course title.
- Development of lesson plans, assignments, and evaluation rubrics for LSC 606, Cataloging and Classification, for the School's NCATE accreditation. 2008
- Howard and Mathilde Rovelstad Scholarship Committee, Chair, 2004-2007
- Assistant Dean Search Committee, Chair, Fall 2007
- Liaison to the Association for Library Collections and Technical Services to bring its preconference program, Cataloging Education and Employer Expectations, to CUA during the 2007 American Library Association Annual Meeting in Washington, D.C.
- Organizer of the colloquium presentation and reception for Tamar Sadeh of Ex Libris on PRIMO June 2007
- Practicum review and design (work with potential supervisors, such as the American Indian Museum internship description revision) 2006-
- Comprehensive examinations (edits, proctoring, and grading), 1990-
- SLIS Web site redesign: Comments and suggestions. Fall 2007
- Conducted surveys of current students and alumni in preparation for the 2005 re-accreditation, 2004-2005
- Technology Committee, 1992-1999 (chair, 1996-1998), 2002-2003 (member)
- Colloquia Committee 1997-1999, 2002-2003.
- Advisor of the CUA Student Chapter of the American Society for Information Science and Technology, 2002-2003
- Visiting Professor Search Committee, 1999, 2000, 2001
- Leader, Participation in the CORC experiment, 1999-2000
- Advisor of the Special Libraries Association Student Chapter, 1993-1999; the group was recognized for outstanding leadership by SLA in 1999.
- COA planning Committee, Task Force on Electronic Presentation of SLIS Reports (team leader) 1997-1998
- COA Planning Committee, Subcommittee on Technology 1996-1998
- NLM practicum coordinator, 1997-1998
- Computer Literacy Workshops: Assisted with the development and evaluation of the workshops, 1996-1998
- Leader, Participation in the InterCat project, 1995-1997

Appendix B – Materials Considered

Exhibit / Appendix	Description
1007	Bateman et al. (2015, October). Natural Language Processing for Extracting Conveyance Graphs, United States Patent Application Publication, no. US 2015/0286630 A1.
1008	Curtotti, M., & McCreath, E. (2010, December). Corpus based classification of text in Australian contracts. In <i>Proceedings of the Australasian Language Technology Association Workshop 2010</i> , pp. 18-26. (“Curtotti”).
1008A	Copy of <i>Curtotti</i> archived by Internet Archive on August 7, 2017, available at https://web.archive.org/web/20170807183515/www.aclweb.org/anthology/U10-1005 .
1008B	ACL Anthology website archived by Internet Archive on Aug. 7, 2011, available at https://web.archive.org/web/20110807051635/http://www.aclweb.org/anthology/. Proceedings of the Australasian Language Technology Association Workshop 2010, archived by Internet Archive on Aug. 3, 2011, available at https://web.archive.org/web/20110803053509/http://aclweb.org/anthology/U/U10/. <i>Curtotti</i> archived on by Internet Archive on July 8, 2011, available at https://web.archive.org/web/20110708111015/http://aclweb.org/anthology/U/U10/U10-1005.pdf.
1008C	Usage records of <i>Curtotti</i> obtained from Google Scholar.

Exhibit / Appendix	Description
1011	Kwak, A., Israelsen, J., Morrison, C., Bambauer, D., & Surdeanu, M. (2022, December). Validity assessment of legal will statements as natural language inference. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6047-6056. (“ <i>Kwak</i> ”).
1011A	Copy of <i>Kwak</i> archived by Internet Archive on February 2, 2023, available at https://web.archive.org/web/20230202142404/https://aclanthology.org/2022.findings-emnlp.447.pdf .
1011B	A pre-publication copy of <i>Kwak</i> on ArXiv, archived by Internet Archive on November 1, 2022, available at https://web.archive.org/web/20221101104607/https://arxiv.org/pdf/210.16989v1.pdf .
1011C	Discovery of the ArXiv copy of <i>Kwak</i> : ArXiv metadata record for <i>Kwak</i> archived by Internet Archive on Nov. 1, 2022, available at https://web.archive.org/web/20221101102929/http://arxiv.org/abs/2210.16989v1 , and the PDF link leads to an archived copy of <i>Kwak</i> , archived on Nov. 1, 2022, available at https://web.archive.org/web/20221101104607/https://arxiv.org/pdf/210.16989v1.pdf .

Exhibit / Appendix	Description
1011D	Discovery of <i>Kwak</i> archived by Internet Archive on February 2, 2023: Metadata record for <i>Kwak</i> at https://aclanthology.org/2022.findings-emnlp.447/ on ACL Anthology website. Wayback Machine calendar for https://aclanthology.org/2022.findings-emnlp.447.pdf. Copy of <i>Kwak</i> archived by Internet Archive on February 2, 2023, available at https://web.archive.org/web/20230202142404/https://aclanthology.org/2022.findings-emnlp.447.pdf.
1011E	Discovery of <i>Kwak</i> archived by Internet Archive on February 2, 2023: ACL Anthology website archived by Internet Archive on March 14, 2023, available at https://web.archive.org/web/20230314133352/https://aclanthology.org/. Conference on Empirical Methods in Natural Language Processing (2022) archived by Internet Archive on March 18, 2023 and a screen capture of the webpage with an entry for <i>Kwak</i>, available at https://web.archive.org/web/20230318065928/https://aclanthology.org/events/emnlp-2022/. Copy of <i>Kwak</i> archived on February 2, 2023, available at https://web.archive.org/web/20230202142404/https://aclanthology.org/2022.findings-emnlp.447.pdf.

Exhibit / Appendix	Description
1011F	Discovery of <i>Kwak</i>, archived by Internet Archive on February 2, 2023: ACL Anthology website archived by Internet Archive on March 14, 2023, available at https://web.archive.org/web/20230314133352/https://aclanthology.org/. Findings of the Association for Computational Linguistics (2022) archived on March 7, 2023 at https://web.archive.org/web/20230307233013/https://aclanthology.org/events/findings-2022/ and a screen capture of the webpage with an entry for <i>Kwak</i>. Copy of <i>Kwak</i> archived on February 2, 2023, available at https://web.archive.org/web/20230202142404/https://aclanthology.org/2022.findings-emnlp.447.pdf.
1011G	Copy of <i>Kwak</i> on SSRN, available at https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID4279728_code575020.pdf?abstractid=4279728&mirid=1.
1011H	Usage Records from SSRN SSRN metadata record for <i>Kwak</i>, archived by Internet Archive on Nov. 28, 2022, available at https://web.archive.org/web/20221128060841/https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4279728. SSRN Author profile for Alice Kwak, archived by Internet Archive on February 2, 2023, available at https://web.archive.org/web/20230202153811/https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=5582596

Exhibit / Appendix	Description
1011I	Cao, Y., Sun, Y., Xu, C., Li, C., Du, J., & Lin, H. (December 6, 2022), CAILIE 1.0: A dataset for Challenge of AI in Law – Information Extraction V1.0, in AI Open, Vol. 3, pages 208-212, available at https://www.sciencedirect.com/science/article/pii/S2666651022000237?via%3Dihub (click “Show more” for received date, revised date, accepted date, available online date, and version of record date).
1011J	Gildea, D., Kan, M., Madnani, N., Teichmann, C., & Villalba, M. (July 20, 2018), The ACL Anthology, Current State and Future Directions, in Proceedings of Workshop for NLP Open Source Software, pages 23-28, available at https://aclanthology.org/W18-2504v1.pdf .

APPENDIX

1008A

← → ↻ web.archive.org/web/20170807183515/www.aclweb.org/anthology/U10-1005 ☆ Sign in

INTERNET ARCHIVE <http://www.aclweb.org/anthology/U10-1005> Go JUL AUG SEP 07 2016 2017 2018 About this capture

Wayback Machine 4 captures 7 Aug 2017 - 18 Apr 2021

1 of 9 Automatic Zoom

Corpus Based Classification of Text in Australian Contracts

Michael Curtotti

Australian National University
Canberra, Australia

michael.curtotti@anu.edu.au

Eric McCreath

Australian National University
Canberra, Australia

eric.mccreath@anu.edu.au

Abstract

Written contracts are a fundamental framework for commercial and cooperative transactions and relationships. Limited research has been published on the application of machine learning and natural

problem affecting the drafting and operation of contracts. The costs of ambiguity in a contract can be enormous. The presence of ambiguity increases the risk of litigation, project failure and loss of commercial relationships. Instances of litigated ambiguity are used in drafting manuals as instructional tales for the unwary drafter (Aitken

Corpus Based Classification of Text in Australian Contracts

Michael Curtotti

Australian National University
Canberra, Australia

michael.curtotti@anu.edu.au

Eric McCreath

Australian National University
Canberra, Australia

eric.mccreath@anu.edu.au

Abstract

Written contracts are a fundamental framework for commercial and cooperative transactions and relationships. Limited research has been published on the application of machine learning and natural language processing (NLP) to contracts. In this paper we report the classification of components of contract texts using machine learning and hand-coded methods. Authors studying a range of domains have found that combining machine learning and rule based approaches increases accuracy of machine learning. We find similar results which suggest the utility of considering leveraging hand coded classification rules for machine learning. We attained an average accuracy of 83.48% on a multi-class labelling task on 20 contracts combining machine learning and rule based approaches, increasing performance over machine learning alone.

1 Introduction

Contracts govern economic transactions and commercial or organisational relationships from trivial purchases to major national infrastructure projects. Any large organisation (whether private or public) must unavoidably invest significant resources in developing and concluding contracts, as the contracts it enters define its legal relations with organisations and individuals with which it interacts. As one author has noted, contracts are an integral part of any business enterprise and it is “difficult to overstate their importance to the business world” (Khouri and Yamouni, 2007, p16). In addition, drafting contracts represents a major economic activity for the legal industry.

Ambiguity is an unresolved problem in natural language processing. It is also an important

problem affecting the drafting and operation of contracts. The costs of ambiguity in a contract can be enormous. The presence of ambiguity increases the risk of litigation, project failure and loss of commercial relationships. Instances of litigated ambiguity are used in drafting manuals as instructional tales for the unwary drafter (Aitken and Butt, 2004, p37).

Our future research goal is to investigate approaches that will help identify such ambiguity, however, a precursor for addressing the problem of ambiguity is accurate labelling or classification of the constituent parts of contract documents.¹ In this paper we report work on classifying ‘lines’ in contracts into 32 classes. We compare three different classification approaches in order to determine which approach may provide the best performance with respect to this task: supervised machine learning; a hand-crafted rule based tagger; or combining hand-crafted rules based tagging with machine learning.

2 Background

Natural language processing involves applying a pipeline of transformations to text. A typical first step is to segment the text into its sentences, after which further processing is applied (Jurafsky and Martin, 2009, p 69). This works well in usual English prose, but is not necessarily well suited to contracts because of the heavy use of layout to embody structure and semantics. In a contract, a sentence may occur within a single paragraph of text, or may be spread over several line breaks due to the use of sub-paragraphing as an aid to comprehension. Such sub-paragraphs may embody disjunctive or conjunctive conditions associated with

¹This task can also be thought of as a document segmentation task, although we here consider it as a classification exercise.

a rule. A ‘clause’² is the typical structural unit which embodies a legal rule or set of rules. Sometimes clauses are organised into sub-clauses and clauses and sub-clauses may be numbered in hierarchical fashion (e.g. 1, 1.1, (a), (i), (A))(Aitken and Butt, 2004, pp 23, 27). Line breaks often occur immediately before data such as headings, clauses, sub-clauses, party names, dates or execution blocks. Being able to label a line as to its character, key content or function thus enriches available information for later NLP. The line is thus a valuable starting point for applying NLP.

3 Related Work

3.1 Classification and Segmentation of Contracts and Legislative Documents

There would appear to be little published work on the application of natural language processing (NLP) or machine learning specifically for the classification of text in legal contracts. An exception to this is (Indukuri and Krishna, 2010) who apply a support vector machine over feature vectors extracted from contract clauses. These feature vectors are comprised of n-grams up to size $n = 4$. They classify contract sentences as clauses or non-clauses and then sub-classify clauses as concerned with payment terms or not. Indukuri and Krishna note the lack of published work on classification of sentences in contract documents for workflow and monitoring purposes.

While also limited, there has been some work on segmentation or classification of legislative corpora using machine learning. (Mencía, 2009) reports 100% precision and recall in machine learning on classification of legislation into articles, sections and parts (although to a task for which he is able to craft an equally accurate regular expression based segmenter). (Bacci et al., 2009) automatically classify plain text legislative documents for the purpose of XML mark up. (Francesconi, 2006), on whose work Bacci et al. build, describes four separate modules for dealing with plain text legacy content which they are seeking to convert to XML tagged text. (Hasan et al., 2008) carry out segmentation of Spanish legislative bulletins using the table of contents as an aid to segmentation. This work, in the legislative field, provides a parallel application domain for the work undertaken here in respect of contracts, although it is

²The term ‘clause’ here refers to a legal rule or rules appearing in a contract, rather than the linguistic entity.

generally true that legislation will tend to conform to far tighter stylistic rules, because of the central control of legislative drafting.

3.2 Combining Machine Learning and Hand Coded Approaches

A number of researchers have combined hand-coded approaches and machine learning to produce better overall results.

(Kipp, 2006) augments handcrafted rules with machine learnt rules for gesture generation.

(McCreath and Kay, 2003) used such a combined approach to improve performance on categorizing emails.

(Park and Zhang, 2003) address the problem of text chunking for the Korean language, combining hand-crafted rules with a memory based machine learning method. A rule based method is used to determine whether an instance may constitute an exceptional case, with machine learning used to correct classifications assigned by the rule based method. They report up to a 2.83% improvement in F-scores over memory-based learning alone.

(Rochery et al., 2002) report the use of hand-crafted rules to enhance the accuracy of machine learning utilising boosting. They apply the method to a multi-labelling classification problem where they are seeking to classify spoken utterances. They attribute increased accuracy to the fact that the rules they develop are not represented in their data at all or do not appear sufficiently to have statistical impact during the training process. They note that the combination has the greatest effect when there are less than 100 examples, with the benefit of the combination decreasing as the number of training examples increases. The method therefore has direct applicability to reducing the data requirement for machine learning (a costly aspect of supervised machine learning).

(Takahashi et al., 2005) utilise a combination of SVM based learning and hand-crafted rules for the task of classifying occupations from social surveys. They effect a combination by using rule derived labels as features for the machine learner. Again the work occurs in a multi-class labelling context (with around 200 occupation codes). They find that SVM approaches are superior to rule-based approaches alone but that combining rule-based and SVM learning produces the best results. They also examine the effect of the number of training examples, noting that the differential ben-

efits of combining methods reduces with increasing data.

In our study we use class labels assigned by a hand-coded tagger as additional features for machine learning to obtain an increase in accuracy in the domain of text classification.

The work reported above, and our own paper, come from diverse domains and illustrate the application of combined rule/ML approaches using a variety of machine learning algorithms.

This suggests the wide utility of using hand-coded rules to improve accuracy of machine learning. Symetrically we may say that machine learning may potentially be used to improve the accuracy of hand-coded rules. Here we essentially follow a hybrid model of development that examines each method separately as well as in combination.

4 Data

4.1 Australian Contract Corpus

The data which forms the basis of this study is a corpus of 256 contracts ('the Australian Contract Corpus') which has been compiled from the web using the Google Australia web search: <http://www.google.com.au>.³ The corpus is constituted of 1043364 tokens (words and punctuation), 42910 sentences, and a vocabulary of 14217. The lexical diversity of this corpus is 73.39.⁴ As far as we are aware there is no published work based on a corpus of Australian contract language, however space does not permit a fuller discussion of the design and analysis of the corpus.

4.2 Data Representation

For the purpose of this study we randomly selected 30 contracts from the Australian Contract Corpus to produce data sets for application of machine learning. The 30 contracts were divided into three

sets: a training set (Set A)⁵, a second set (Set B)⁶ (primarily used as a test set, although sometimes combined with Set A to form a larger training set) and an additional test set (Set C)⁷, each of 10 contracts. Set C was added to increase both available data and the number of contracts on which testing could be carried out.

To simplify processing all contracts were converted into text files as a preprocessing step.⁸ Also we removed material that typically does not appear in contracts (primarily guideline notes for drafters).

For purposes of our study each 'line' or 'paragraph' in a contract constituted a data point. Sets A, B and C provided 1825, 2157 and 2231 lines/data points, respectively.

To classify our data for machine learning we first employed our hand-coded tagger to produce a labelled data set constituted of a line and primary and secondary labels according to our classification system (see below). The classifications applied by the hand-coded tagger were then manually corrected to remove any labelling errors. Both for the reason that hand tagging data from scratch is more error prone than correcting previously machine tagged data, and because of the high labour in manual tagging, machine assisted tagging is a standard method used in corpus creation (Biber et al., 1998, p 262), ((ed) Christopher S. Butler, 1992, p 131).

4.3 Feature Selection

Features were then extracted from our labelled lines using a hand-coded feature extractor to extract up to 40 features to be used for machine learning, with the primary class label earlier applied to each line serving as the target for machine learning.⁹ Features for machine learning were selected based on assumed relevance to the intended classification.¹⁰ While a bag of word representa-

³The search phrase we employed was: "clause party agreement", limiting the search to "pages from Australia" and the filetype to ".doc". The collection of the corpus was undertaken in the period 6 - 24 December 2009. Each document was visually inspected by one of the authors to verify that it was an Australian contract and documents were added to the corpus in order of their appearance in Google search results. We have made available the list of URLs of the documents that make up the corpus at: <http://cs.anu.edu.au/people/Michael.Curtotti/>. We also make code and data referred to in this paper available at the same location.

⁴Lexical diversity refers to the ratio of the total number of tokens in the corpus to the total number of types of tokens i.e. tokens/vocabulary. The statistics and information provided here were extracted using the NLTK. (Bird et al., 2009).

⁵Contracts numbered in our corpus: 55, 74, 77, 91, 94, 144, 174, 185, 208 and 213

⁶Contracts 11, 12, 164, 193, 196, 199, 249, 59, 64, 9

⁷Contracts 200, 254, 175, 180, 120, 102, 207, 127, 75 and 79

⁸Notably, this step does have the downside of discarding valuable style and layout information found in the word document format.

⁹In some instances the assignment of class is not entirely disjunctive and implies a priority ordering of potential classifications. For instance a clause may contain address details, and be classified as such rather than as a clause. The feature extractor removed secondary labels.

¹⁰The full list of features is as follows: 1. relative position of a line within the contract; 2. line end punctuation; 3. num-

tion with automatic feature extraction was trialed, this was found to result in lower accuracy than a manually selected (but small) number of features. Traditionally, ‘feature engineering’ or ‘selection’ in text classification has focussed on ‘bag of words’ or proximate features such as n-grams or noun phrases (Scott and Matwin, 1999). Our selection of features might be regarded as ‘hybrid’ in that they implicitly encode domain knowledge and ‘rules’. For example, we expect a definition line to contain the widely used word ‘means’ as in ‘x means y’. We would expect clause headings to display a different pattern of parts of speech occurrence, as compared to clauses themselves. We also consider features such as the relationship between the particular line/datapoint and its neighbouring lines (e.g. is the preceding line long or short), or the document as a whole (e.g. relative placement of the line within the document).

Features used in the related work carried out by others on contracts and legislation include: n-grams (Indukuri), bag of word together with capitalisation and character patterns (Mencia), the use of sequential word and token data as a HMM to predict state where state is the classification of the portion of text (Bacci/Francesconi) and the use of word terms identifying titles and lists of indexed terms (Hasan).

4.4 Classes

We adopted 32 classes for our classification scheme. These classes represented significant

being at the beginning of the line; 4. line length in characters; 5. line length in tokens; 6. number of nouns; 7. number of adjectives; 8. number of verbs; 9. number of prepositions; 10. number of coordinating conjunctions and cardinal numbers; 11. number of modal verbs; 12. number of personal pronouns, possessives; 13. number of adjectives; 14. whether must occurs; 15. whether may occurs; 16. whether shall occurs; 17. whether means occurs; 18. the position of the word means in the line; 19. whether the word ‘include’ occurs; 20. the position of the word ‘includes’ in the line; 21. whether the phrase ‘has the same meaning’ occurs; 22. whether the line begins with a capital letter followed by a stop or a space; 23. do the letters ‘abn’ or ‘acn’ appear; 24. does the word address appear near the beginning of the line; 25. does the word ‘contact’ appear near the beginning of the line; 26. does the word ‘email’ appear near the beginning of the line; 27. does the word ‘fax’ appear near the beginning of the line; 28. does the word ‘note’ appear at the beginning of the line; 29. does the word ‘phone’ appear at the beginning of the line; 30. do the terms ‘web’ or ‘www’ appear near the beginning of the line; 31-36. token lengths of 3 lines before and after the data line; 37. the tag applied by the hand coded tagger to the data line; 38. whether the contract from which the line comes has clause headings; 39. whether the contract from which the line comes has clause sub-headings; 40. whether the contract has a schedule.

structural elements (such as clause headings or content lines) or sometimes key contract meta data (for instance the parties, the date on which an agreement is made, email addresses, ABNs etc).¹¹ Of particular interest to us, was an ability to accurately identify clausematter (the primary location of legal rules of a contract) and clause headings (which in some contracts effectively mark boundaries between major rule sets).

5 Experimental Evaluation

Analysis of our data sets was carried out using two major methods: analysis using the training Set A measuring performance on test Set B (i.e. testing for accuracy of classification of lines in Set B); and analysis using training Set A where each contract in test Sets B and C was used individually as a test set (testing for accuracy of classification of lines of each individual contract). The former testing approach is more widely used but the latter is more consistent with our intended end application: that is, developing an ability to accurately classify lines in a previously unseen contract as an individual document.

Set A consisted of 1825 lines of data, Set B consisted of 2157 lines of data and Set C consisted of 2231 lines of data.

5.1 Tools

The following tools were used to carry out machine learning tasks reported in this paper:

1. the *python* programming language was used to develop a hand coded line tagger, a feature extractor and code for evaluation of the performance of the hand coded tagger;
2. the Natural Language Toolkit (NLTK)(Bird et al., 2009) was used in corpus development and in the hand coded tagger for application of parts of speech tags; and
3. the Weka data mining software(Hall et al., 2009) was used to carry out machine learning.

¹¹The following class labels were applied: CLAUSE-MATTER, CLAUSEHEAD, DEF, CONTENTSHEAD, BLANK, DROSS, AND, BETWEEN, DATEMADE-LINE, PARTIESHEAD, PARTYLINE, RECITALHEAD, RECITALLINE, HEAD, CONTENTLINE, PRELIM, OPERATIVEPROVISIONLINE, NUMBEREDLINE, SCHEDULEHEAD, SCHEDULEITEM, EXECUTIONBLOCK, NOTE, ABNLINE, ADDRESSLINE, EMAILLINE, WEBSITE, FAXLINE, PH-LINE, REF-LINE, SCH-OTHERMATTER, CONTACTOFFICER, TITLELINE.

5.2 Hand-Coded Tagger

Our hand-coded tagger follows a pipeline methodology. A contract is separated into a list of lines. In the first phase, each line is passed through a series of “if-then-else” routines (essentially utilising regular expressions) to assign primary and secondary labels to the line. This produces a contract with enhanced information content, in the form of a tuple of labels and line text. This output is then fed through a series of methods, progressively bootstrapping improvements to labelling, based on the information available when the method is applied. For instance content lines will be identified, these are then used in later methods to seek to identify clause headings. In addition, the tagger extracts contract meta data such as: a list of definitions, a list of clause headings, where the beginning of operative provisions occurs, where the schedule begins. Such features may or may not occur in a contract. In addition the tagger also provided a method for hand correcting tagging which was used to create our data set for supervised machine learning purposes.

Development of the tagger relied heavily on domain knowledge to identify logical tests for label application. For instance the occurrence of numbering at the beginning a line was used to identify clause headings, or the presence of a common text pattern which occur in definitions i.e. (the pattern *[“word[1-3]] means [...]”* was used as a criterion for identifying a line as a definition).

	$\bar{x}$	s
On Set B Contracts	86.27%	NA
Contract x Contract	82.84%	10.85%

Table 1: Accuracy of Line Tagger.

The first row of Table 1 reports accuracy of the hand coded tagger when applied to Set B as the test set (the % accuracy of tagging of individual lines in Set B is reported). The second row of Table 1 reports average accuracy with which the lines in each of 20 contracts in Sets B and C are tagged (standard deviation is also reported). The high variance of the results at a contract by contract level suggests a need for future work to identify if issues such as contract ‘type’ may account for the differences (the lowest level of accuracy was around 50% and the highest around 95%).

Table 2 provides measures of precision, recall and F-measure attained by the hand coded tag-

	Precision	Recall	F-Measure
CL.HEAD	0.98	0.83	0.89
CL.MATTER	0.91	0.91	0.91
DEF	1.00	0.66	0.80

Table 2: Performance results for Hand-Coded Line Tagger

ger on Set B for key data items: clause headings, clauses themselves and definitions.¹² These results on key measures are sufficient for our initial purposes and will allow us to focus subsequent work on addressing ambiguity in the substantive rules found in a contract.

5.3 Supervised Machine Learning

Column 1 of Table 3 reports the accuracy of various supervised machine learning algorithms with Set A as training set and Set B as test set. Of the algorithms applied, Random Forest (implemented using 100 trees and 30 random features) performed best. No ML method performed better than the hand coded tagger tested on the same test set (see row 1 of Table 1).

5.4 Machine Learning vs Hand-coded methods

Given our ultimate focus on developing applied tools, we are primarily concerned to identify the best methods to maximise the accuracy of our classification task. Whether this proves to be machine learning methods or hand-coded methods or a combination of the two is immaterial from this viewpoint (other than in regard of development costs).

Also it is useful to consider whether methods and insights from one method may assist in the other. For instance, hand-coded rules that successfully labelled lines directly suggested ‘features’ to be extracted for machine learning purposes. Conversely the identification of ‘features’ for machine learning can be regarded as a programming abstraction where the algorithm is ‘under the hood’ but human intervention is required in the form of feature selection for ‘learning’ to occur (implicitly encoding of rules in the feature set). In deriving our results we used a confusion matrix as a tool to assist in identifying areas where the hand-

¹²Precision is $\frac{TP}{TP+FP}$, Recall is $\frac{TP}{TP+FN}$, and F-Measure is $\frac{2PR}{P+R}$, where TP is true positives, FP is false positives, FN is false negatives, P is precision, and R is recall.

coded tagger required improvement and as a common baseline for performance. In certain forms of machine learning, the resulting machine learning model can be represented as code in “if, then, else” form (as in the output of a decision list or decision tree learner). We may thus conceive of machine learning and hand coding of rules as different ‘programming paradigms’ - a perspective that encourages us to apply insights across the boundary between them.

Further, we wished to explore the relative accuracy of machine learning vs hand coding for this particular task. Such an exploration is valuable from a cost benefit point of view. Supervised machine learning (where we have focused our attention) requires the production of human annotated data (which is costly to produce) as well as feature testing and refinement. Also machine learning may be computationally intensive. Production of an accurate hand-coded tagger requires expert knowledge and considerable testing and code refinement in order to achieve a high level of accuracy (which again is resource intensive). We did not undertake quantitative data collection in respect of time employed in the two methods, particularly as we followed an iterative development model that switched between rule based and machine learning focussed work and as we were searching for optimal outcomes. Nonetheless we anecdotally report that each method presented significant demands in terms of development time.

In the results reported above we note that no machine learning approach alone exceeded the performance of the hand coded tagger when considering performance on test Set B as a whole. This is the result also found at contract by contract level (see below Table 4).

5.5 Leveraging Hand-Coded Labels for Machine Learning

As noted above we also wished to explore whether utilising the output from the hand coded tagger as input features for machine learning might improve the accuracy of machine learning alone. A converse question that could be framed is whether the rule based tagger’s performance could be improved by combination with machine learning.

We carried out two trials that bear on these questions:

1. machine learning alone and machine learning enhanced with rule based input on an entire

test set (Set B); and

2. machine learning alone, rule alone and machine learning enhanced with rule based input at contract by contract level.

ML Algorithm	ML Alone	ML + Rule
Naive Bayes	82.38%	84.75%
SMO (SVM)	82.80%	85.40%
Cl. via Regression	83.91%	87.07%
Decision Tree	82.01%	87.02%
Bagging	83.54%	87.76%
Majority Vote	84.42%	87.11%
Random Forest	85.12%	86.69%

Table 3: Comparative Performance ML alone and ML + Rule on train and test set

Column 2 of Table 3 shows the improvement in performance of various machine learning methods when tags applied by the hand coded tagger were provided as input features. All methods show improvement with addition of the rule derived feature on Set B as a whole.¹³

We report below a comparison of accuracy of machine learning, rule based tagging and a combination of both in respect of Decision Trees. Decision trees offer the advantage that the output is easily interpretable as a rule set for tagging.

Method	$\bar{x}$	s
ML Alone	79.12%	10.71%
Rule Alone	82.84%	10.85%
ML + Rule	83.48%	9.83%

Table 4: Tagging Accuracy Mean and Standard Deviation - Contract by Contract

Although the results reported in Table 4 are different (as different contracts are included in the tested contracts)¹⁴ average accuracy continued to show a differential between machine learning alone, and machine learning leveraged with hand rules, as described above. On this combination ML + Rule also outperformed the rule based tagger alone, but by a very narrow margin. In 17 of 20 cases ML + Rule improved or did not worsen performance of ML alone. In 15 of 20 cases, ML +

¹³The training set for the results reported in Table 3 was Set A. The test set was Set B.

¹⁴The training set was Set A. Sets B and C provided 20 individual contracts (effectively 20 test sets) for testing.

rule was no worse or outperformed the hand coded tagger alone. Notably variation is quite high as compared to the mean. Testing for statistical significance we can reject a null hypothesis that the mean accuracy for ML + Rule is the same as for ML alone. We cannot reject a null hypothesis that ML + Rule is the same as Rule alone.¹⁵ We note also that this investigation was only carried out in respect of the decision tree learner.

5.5.1 Is it more effective to increase data or use rule based features?

To further explore the question of how much effort might be saved by combining rule based approaches with machine learning alone we tested the effect of adding random noise to the feature set provided by the hand coded tagger. We found that we had to randomly flip over 45% of these tags before they ceased to impact positively on classification accuracy. This suggests that even poorly prepared hand coded rules can improve performance.¹⁶

We also tested the effect of incrementally increasing the amount of available training data to explore any decrease in differential in performance. Testing using Set C as the test set and taking training data from Sets A and B, we progressively increased training data by increments of 500 instances of data. Differential performance decreases progressively, starting around 5% and reducing to 2.5% as more training data is made available.¹⁷ (See Table 5.) Extrapolating from these figures we would expect at a minimum that training data would have to be doubled to remove differential accuracy. Relatively little effort in developing rule based features may substitute for considerable work in creating additional supervised data.

¹⁵We gratefully acknowledge the assistance of Bob Forrester of the ANU Statistical Consulting Unit in deriving this result, undertaken utilising a generalised linear model to undertake model fitting adjusting for different file (contract) types. Predictions of estimated mean proportions were: ML Alone 0.7760 (s.e. 0.007181), Rule alone 0.8106 (s.e. 0.006775), ML+ Rule 0.8181 (s.e. 0.006677).

¹⁶Using a Decision Tree, ML + Rule accuracy on test Set B with training Set A, was 87.0% with zero noise, 86.9% with 5% noise, 84.7% with 10% noise, 85.2% with 20% noise, 84.4% with 30% noise, 83.9% with 40% noise, 82.8% with 45% noise, 82.4% with 50% noise. Decision Tree accuracy on ML alone on this training and test set was 82.0%.

¹⁷The data has a negative correlation of -0.70

Data	0.5	1K	1.5	2K	2.5	3K	3.5	4K
Diff.	5.1	3.4	5	4.2	3.3	3.1	3.9	2.6

Table 5: Difference in % accuracy of ML alone and ML with rule based feature with increasing training data.

5.5.2 Decision tree output

An examination of the decision tree produced after adding the output of the hand-coded tagger to machine learning is of interest.

```

LTlabel = #CLAUSEMATTER#
| tknLngth <= 4
|| nounNum <= 1: #CLAUSEMATTER# (6.0/2.0)
|| nounNum > 1
||| linePos. <= 0.383929: #PRELIM# (2.0)
||| linePos. > 0.383929: #CLAUSEHEAD# (2.0)
| tknLngth > 4: #CLAUSEMATTER# (508.0/7.0)
: :
: :
LTlabel = #PH-LINE#
| prepNum <= 1: #PH-LINE# (5.0)
| prepNum > 1: #CLAUSEMATTER# (2.0)
LTlabel = #EMAILLINE#: #EMAILLINE# (3.0)
LTlabel = #ABNLINE#
| preLine2 <= 0: #TITLELINE# (2.0/1.0)
| preLine2 > 0: #ABNLINE# (3.0)

```

Figure 1: Part of the decision tree learnt when the labels from the hand-coded are provided as an attribute.

Using only the machine learning feature set, the size of the pruned tree is 199 with 108 leaves. After addition of labels provided by our hand-coded tagger the size of the decision tree is reduced to 105 with 69 leaves. Figure 1 shows part of this decision tree and as illustrated, the decision tree generally begins with the label assigned by the hand-coded tagger 'LTlabel' and amends the label (where necessary) utilizing other features available to the machine learner. Some labels (e.g. EMAILLINE) were simply adopted without change. In the case of the CLAUSEMATTER label, the machine learner used the number of tokens in the line ('tknLngth'), number of nouns in the line ('nounNum') and the relative position of the line in the contract ('linePos.') to re-assign the class of the line. The significance of the use of labels assigned by the hand-coded tagger is that the decision-tree identifies that attribute as providing the most significant information gain (that is reduction in entropy) and therefore partitions on that feature (Callan, 2003, p245). This remains true in this example irrespective of the class to which

the data is being assigned.¹⁸ It is obvious that the best feature for assigning the correct label for an item of data is the correct or probably correct label (if one can obtain it without undue endeavour). From a practical viewpoint some effort directing to hand-crafting code to extract appropriate class labels may thus assist machine learning.

6 Conclusions and Future Work

In terms of the overall performance of the methods described in this paper the closest point of comparison is the work reported by Indukuri and Krishna (Indukuri and Krishna, 2010). While comparisons using different methods, software and data are of dubious validity, we may note their broad similarity. With a training set of 10 contracts and a test set of 20 contracts, the contract-by-contract average F-measure for the clausematter class was 0.8632. Also in 14 contracts of the 20 tested the F-measure for the clausematter class was above 0.85. With a training set of 10 contracts (Set A) and a test set of 10 contracts (Set B) (as a single test set) the F-measure for the clausematter class was 0.921.¹⁹ Using 4-grams as features, Indukuri and Krishna report an accuracy of 83.48% on the task of classification clauses from non-clauses, applied to 73 sentences to be so classified. Whereas the task in that case was a binary classification task over one contract, we report results on multi-class classification over test sets of up to 20 contracts.²⁰

The work we have undertaken and related work reviewed in this paper (Section 3) suggests the potentially broad utility of combining rule based and machine learning methods. One potential hybrid classification method in its simplest formulation may look something like this:

1. determine the required or anticipated level of accuracy for the intended application;
2. develop a simple code rule set for the classification task (i.e. with minimal expenditure of resources);
3. if classification accuracy is sufficient machine learning would not be required, if not

¹⁸The Weka decision tree implementation, “J48”, is based on the Quinlan’s C4.5 (Quinlan, 1986).

¹⁹The results reported here are derived using a random forest algorithm and using the hand coded tag as an input feature for classification.

²⁰Coincidentally the 83.48% figure is the same as our average result over all classes for 20 contracts.

proceed as usual in development of machine learning using the output of hand coded rules as an input to machine learning.

Such a development model does not avoid the development of a supervised training set, as such a set is required both to assess the accuracy of the rule based tagger as well as the machine learner (should development proceed to that stage). However it may reduce the amount of data required to attain a desired level of accuracy. The method could of course be applied iteratively, if the accuracy level is insufficient after a first iteration.

Future work to be explored includes the effective “typing” of contracts (particularly in light of the high variance of performance on individual contracts). For instance, some contracts use clause headings, others do not. Some use schedules, others do not, etc. Such type information may further assist in the classification task (the accuracy of which varies considerably depending on the contract to which it is applied). In typical preprocessing, documents in corpora are converted to plain text. This results in the loss of layout and hierarchical information found in word documents. The preservation of such information for use in classification tasks may improve accuracy.

The preliminary work here is also a precursor to work on identification of ambiguity and the development of practical tools to assist in contract drafting. In separate work, we intend also to describe and analyse the contract corpus that is referred to here as the basis of this work.

References

- J. K. Aitken and P. Butt. 2004. *Piesse The Elements of Drafting*. Lawbook Co, 10th edition.
- L. Bacci, P. Spinosa, C. Marchetti, R. Battistoni, I. Florence, I. Senate, and I. Rome. 2009. Automatic mark-up of legislative documents and its application to parallel text generation. In *Proc. of LOAIT Workshop*, pages 45–54.
- D. Biber, S. Conrad, and R. Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.
- Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O’Reilly Media Inc.
- Rob Callan. 2003. *Artificial Intelligence*. Palgrave Macmillan.

- (ed) Christopher S. Butler. 1992. *Computers and written texts*. Blackwell.
- E. Francesconi. 2006. The Norme in Rete Project: Standards and Tools for Italian Legislation. *International Journal Legal Information*, 34:358.
- Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software. *SIGKDD Explorations*, 11(1).
- I. Hasan, J. Parapar, and R. Blanco. 2008. Segmentation of legislative documents using a domain-specific lexicon. In *Proceedings of the 19th International Conference on Database and Expert Systems Application*, pages 665–669.
- Kishore Varma Indukuri and P. Radha Krishna. 2010. Mining e-contract documents to classify clauses. In *COMPUTE '10: Proceedings of the Third Annual ACM Bangalore Conference*, pages 1–5, New York, NY, USA. ACM.
- Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, 2nd edition.
- D. Khoury and Y. S. Yamouni. 2007. *Understanding Contract Law*. Lexis Nexis Butterworths, 7th edition.
- M. Kipp. 2006. Creativity meets automation: Combining nonverbal action authoring with rules and machine learning. In *Intelligent Virtual Agents*, pages 230–242. Springer.
- Eric McCreath and Judy Kay. 2003. Iems: Helping Users Manage Email. *User Modeling 2003*, pages 146–146.
- Eneldo Loza Mencía. 2009. Segmentation of legal documents. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 88–97. ACM.
- Seong-Bae Park and Byoung-Tak Zhang. 2003. Text chunking by combining hand-crafted rules and memory-based learning. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 497–504.
- J.R. Quinlan. 1986. Induction of decision trees. *Machine Learning*, 1:81–106.
- M. Rochery, R. Schapire, M. Rahim, N. Gupta, G. Ricciardi, S. Bangalore, H. Alshawi, and S. Douglas. 2002. Combining prior knowledge and boosting for call classification in spoken language dialogue. In *Proceedings of the IEEE International Conference of Acoustics and Speech and Signal Processing*.
- S. Scott and S. Matwin. 1999. Feature Engineering for Text Classification. In *Machine learning: proceedings of the Sixteenth International Conference (ICML'99), Bled, Slovenia, June 27-30, 1999*, page 379. Morgan Kaufmann Pub.
- K. Takahashi, H. Takamura, and M. Okumura. 2005. Automatic occupation coding with combination of machine learning and hand-crafted rules. *Advances in Knowledge Discovery and Data Mining*, pages 269–279.

APPENDIX

1008B

← → ↺ web.archive.org/web/20110807051635/http://www.aclweb.org/anthology/ ☆ Sign in

INTERNET ARCHIVE WayBackMachine http://www.aclweb.org/anthology/ 1,180 captures 13 Dec 2002 – 22 Dec 2011

JUL AUG SEP 07 2010 2011 2012 About this capture

ACL Anthology

A Digital Archive of Research Papers in Computational Linguistics

Search the Anthology via Google **NEW** via Searchbench @ DFKI **NEW** via AAN @ UMich

The ACL Anthology currently hosts over 19,900 papers on the study of computational linguistics. [Subscribe to the mailing list](#) to receive announcements and updates to the Anthology.

NEW The Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing are now listed in the ACL Anthology, as well as its seven workshops.

ACL events

Journal: **Intro** **FS MT&CL** 74-79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 00 01 02 03 04 05 06 07 08 09 10 11
ACL: **Intro** 79 80 81 82 83 84* 85 86 87 88 89 90 91 92 93 94 95 96 97* 98* 99 00 01 02 03 04 05 06* 07 08* 09* 10 11
EACL: **Intro** 83 85 87 89 91 93 95 97* 99 03 06 09
NAACL: **Intro** 00* 01 03 04 06* 07* 09* 10*
SemEval: 98 01 04 07 10
ANLP: **Intro** 83 88 92 94 97 00*
EMNLP: 01 02 03 04 05 06 07* 08 09 10 **NEW** 11
Workshops: 90 91 93 94 95 96 97 98 99 00 01 02 03 04 05 06 07 08 09 10 **NEW** 11
SIGs: ANN BIOMED **UPDATED** DAT DIAL FSM **UPDATED** GEN HAN **UPDATED** LEX MEDIA MOL **UPDATED** MT NLL PARSE MORPHON **UPDATED** SEM SEMITIC WAC

Other Events

COLING: 65 67 69 73 80 82 84* 86 88 90 92 94 96 98* 00 02 04 06* 08 10
HLT: 86 89 90 91 92 93 94 01 03* 04* 05 06* 07* 08* 09* 10*
IJCNLP: 05 08 09*
LREC: 08
PACLIC 95 96 98 99 00 01 02 03 04 05 06 07
Rocling Intro 88 89 90 91 92 93 94 95 96 97 98 99 00 01 02 03 04 05 06 07 08 09 10
TINLAP: 75 78 87
Donors Needed: COLING-65, any missing COLING

ALTA Intro 03 04 05 06 07 08 09 10
RANLP 09
MUC: 91 92 93 95 98
Tipster: 93 96 98
In Progress: Finite String, SensEval 2001

The Wayback Machine - <https://web.archive.org/web/20110807051635/http://www.aclweb.org/anthology/>



ACL Anthology

A Digital Archive of Research Papers in Computational Linguistics

Search the Anthology

via Google

NEW

via Searchbench @ DFKI

NEW

via AAN @ UMich

The ACL Anthology currently hosts over 19,900 papers on the study of computational linguistics. [Subscribe to the mailing list](#) to receive announcements and updates to the Anthology.

NEW The [Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing](#) are now listed in the ACL Anthology, as well as [its seven workshops](#).

ACL events

Journal: [Intro](#) [FS](#) [MT&CL](#) [74-79](#) [80](#) [81](#) [82](#) [83](#) [84](#) [85](#) [86](#) [87](#) [88](#) [89](#) [90](#) [91](#) [92](#) [93](#) [94](#) [95](#) [96](#) [97](#) [98](#) [99](#) [00](#) [01](#) [02](#) [03](#) [04](#) [05](#) [06](#) [07](#) [08](#) [09](#) [10](#) [11](#)

ACL: [Intro](#) [79](#) [80](#) [81](#) [82](#) [83](#) [84](#)* [85](#) [86](#) [87](#) [88](#) [89](#) [90](#) [91](#) [92](#) [93](#) [94](#) [95](#) [96](#) [97](#)* [98](#)* [99](#) [00](#) [01](#) [02](#) [03](#) [04](#) [05](#) [06](#)* [07](#) [08](#)* [09](#)* [10](#) [11](#)

EACL: [Intro](#) [83](#) [85](#) [87](#) [89](#) [91](#) [93](#) [95](#) [97](#)* [99](#) [03](#) [06](#) [09](#)

NAACL: [Intro](#) [00](#)* [01](#) [03](#) [04](#) [06](#)* [07](#)* [09](#)* [10](#)*

SemEval: [98](#) [01](#) [04](#) [07](#) [10](#)

ANLP: [Intro](#) [83](#) [88](#) [92](#) [94](#) [97](#) [00](#)*

EMNLP: [01](#) [02](#) [03](#) [04](#) [05](#) [06](#) [07](#)* [08](#) [09](#) [10](#) NEW [11](#)

Workshops: [90](#) [91](#) [93](#) [94](#) [95](#) [96](#) [97](#) [98](#) [99](#) [00](#) [01](#) [02](#) [03](#) [04](#) [05](#) [06](#) [07](#) [08](#) [09](#) [10](#) NEW [11](#)

SIGs: [ANN](#) [BIOMED](#) UPDATED [DAT](#) [DIAL](#) [FSM](#) UPDATED [GEN](#) [HAN](#) UPDATED [LEX](#) [MEDIA](#) [MOL](#) UPDATED [MT](#) [NLL](#) [PARSE](#) [MORPHON](#) UPDATED [SEM](#) [SEMITIC](#) [WAC](#)

Other Events

COLING: [65](#) [67](#) [69](#) [73](#) [80](#) [82](#) [84](#)* [86](#) [88](#) [90](#) [92](#) [94](#) [96](#) [98](#)* [00](#) [02](#) [04](#) [06](#)* [08](#) [10](#)

HLT: [86](#) [89](#) [90](#) [91](#) [92](#) [93](#) [94](#) [01](#) [03](#)* [04](#)* [05](#) [06](#)* [07](#)* [08](#)* [09](#)* [10](#)*

IJCNLP: [05](#) [08](#) [09](#)*

LREC: [08](#)

PACLIC [95](#) [96](#) [98](#) [99](#) [00](#) [01](#) [02](#) [03](#) [04](#) [05](#) [06](#) [07](#)

Rocling [Intro](#) [88](#) [89](#) [90](#) [91](#) [92](#) [93](#) [94](#) [95](#) [96](#) [97](#) [98](#) [99](#) [00](#) [01](#) [02](#) [03](#) [04](#) [05](#) [06](#) [07](#) [08](#) [09](#) [10](#)

TINLAP: [75](#) [78](#) [87](#)

Donors Needed: [COLING-65](#), any missing COLING

ALTA [Intro](#) [03](#) [04](#) [05](#) [06](#) [07](#) [08](#) [09](#) [10](#)

RANLP [09](#)

MUC: [91](#) [92](#) [93](#) [95](#) [98](#)

Tipster: [93](#) [96](#) [98](#)

In Finite String, SensEval
Progress: 2001

*: denotes a joint meeting

» [Toggle Notes](#)

Join the Association for Computational Linguistics (ACL): Benefits include discounts on conferences and publications, and membership in special interest groups.

» [Toggle Copyright and Credits](#)

[Min-Yen Kan](#) (Editor, 2008-) / [Steven Bird](#) (Editor, 2001-2007)

← → ↻ web.archive.org/web/20110803053509/http://aclweb.org/anthology/U/U10/ ☆ Sign in

INTERNET ARCHIVE Go 47 captures 2 Jul 2011 – 30 Mar 2019

waybackmachine

JUL AUG SEP 03 2010 2011 2012 About this capture

 **ACL Anthology**
A Digital Archive of Research Papers in Computational Linguistics

Google Custom Search Search this volume

Proceedings of the Australasian Language Technology Association Workshop 2010

[U10-1](#) [bib]: **Entire volume**

[U10-1000](#): *Front Matter*

[U10-1001](#) [bib]: **Rodolfo Delmonte**
Opinion Mining, Subjectivity and Factuality

[U10-1002](#) [bib]: **Casey Whitelaw**
Language Technology: A View From The Trenches

[U10-1003](#) [bib]: **Timothy Baldwin; Marco Lui**
Multilingual Language Identification: ALTW 2010 Shared Task Data

[U10-1004](#) [bib]: **Shunichi Ishihara**
Variability and Consistency in the Idiosyncratic Selection of Fillers in Japanese Monologues: Gender Differences

[U10-1005](#) [bib]: **Michael Curtotti; Eric McCreath**
Corpus Based Classification of Text in Australian Contracts

[U10-1006](#) [bib]: **Li Wang; Su Nam Kim; Timothy Baldwin**
Thread-level Analysis over Technical User Forum Data

[U10-1007](#) [bib]: **Susan Howlett; Mark Dras**
Dual-Path Phrase-Based Statistical Machine Translation

The Wayback Machine - <https://web.archive.org/web/20110803053509/http://aclweb.org:80/anthology/U/U10/>



ACL Anthology

A Digital Archive of Research Papers in Computational Linguistics



Custom Search

Search this volume

Proceedings of the Australasian Language Technology Association Workshop 2010

[U10-1](#) [\[bib\]](#): **Entire volume**

[U10-1000](#): *Front Matter*

[U10-1001](#) [\[bib\]](#): **Rodolfo Delmonte**
Opinion Mining, Subjectivity and Factuality

[U10-1002](#) [\[bib\]](#): **Casey Whitelaw**
Language Technology: A View From The Trenches

[U10-1003](#) [\[bib\]](#): **Timothy Baldwin; Marco Lui**
Multilingual Language Identification: ALTW 2010 Shared Task Data

[U10-1004](#) [\[bib\]](#): **Shunichi Ishihara**
Variability and Consistency in the Idiosyncratic Selection of Fillers in Japanese Monologues: Gender Differences

[U10-1005](#) [\[bib\]](#): **Michael Curtotti; Eric McCreath**
Corpus Based Classification of Text in Australian Contracts

[U10-1006](#) [\[bib\]](#): **Li Wang; Su Nam Kim; Timothy Baldwin**
Thread-level Analysis over Technical User Forum Data

[U10-1007](#) [\[bib\]](#): **Susan Howlett; Mark Dras**
Dual-Path Phrase-Based Statistical Machine Translation

[U10-1008](#) [\[bib\]](#): **Yue Li; David Martinez**
Information Extraction of Multiple Categories from Pathology Reports

[U10-1009](#) [\[bib\]](#): **Marco Lui; Timothy Baldwin**
Classifying User Forum Participants: Separating the Gurus from the Hacks, and Other Tales of the Internet

[U10-1010](#) [\[bib\]](#): **Alexandra L. Uitdenbogerd**
Fun with Filtering French

[U10-1011](#) [\[bib\]](#): **Sze-Meng Jojo Wong; Mark Dras**
Parser Features for Sentence Grammaticality Classification

[U10-1012](#) [\[bib\]](#): **Diego Molla**
A Corpus for Evidence Based Medicine Summarisation

[U10-1013](#) [\[bib\]](#): **Jette Viethen; Robert Dale**
Speaker-Dependent Variation in Content Selection for Referring Expression Generation

[U10-1014](#) [\[bib\]](#): **Dominick Ng; Matthew Honnibal; James R. Curran**
Reranking a wide-coverage ccg parser

[U10-1015](#) [\[bib\]](#): **Simon Zwarts; Mark Johnson; Robert Dale**
Repurposing Corpora for Speech Repair Detection: Two Experiments

[Back to the ACL Anthology Home](#)

← → ↻ web.archive.org/web/20110708111015/http://aclweb.org/anthology/U/U10/U10-1005.pdf ☆ Sign in

INTERNET ARCHIVE WayBackMachine http://aclweb.org/anthology/U/U10/U10-1005.pdf Go JUN JUL AUG 08 2010 2011 2012 About this capture

1 of 9 Automatic Zoom

Corpus Based Classification of Text in Australian Contracts

Michael Curtotti

Australian National University
Canberra, Australia

michael.curtotti@anu.edu.au

Eric McCreath

Australian National University
Canberra, Australia

eric.mccreath@anu.edu.au

Abstract

Written contracts are a fundamental framework for commercial and cooperative transactions and relationships. Limited research has been published on the application of machine learning and natural language processing (NLP) to contracts. In this paper we report the classification

problem affecting the drafting and operation of contracts. The costs of ambiguity in a contract can be enormous. The presence of ambiguity increases the risk of litigation, project failure and loss of commercial relationships. Instances of litigated ambiguity are used in drafting manuals as instructional tales for the unwary drafter (Aitken and Butt, 2004, p37).

Corpus Based Classification of Text in Australian Contracts

Michael Curtotti

Australian National University
Canberra, Australia

michael.curtotti@anu.edu.au

Eric McCreath

Australian National University
Canberra, Australia

eric.mccreath@anu.edu.au

Abstract

Written contracts are a fundamental framework for commercial and cooperative transactions and relationships. Limited research has been published on the application of machine learning and natural language processing (NLP) to contracts. In this paper we report the classification of components of contract texts using machine learning and hand-coded methods. Authors studying a range of domains have found that combining machine learning and rule based approaches increases accuracy of machine learning. We find similar results which suggest the utility of considering leveraging hand coded classification rules for machine learning. We attained an average accuracy of 83.48% on a multi-class labelling task on 20 contracts combining machine learning and rule based approaches, increasing performance over machine learning alone.

1 Introduction

Contracts govern economic transactions and commercial or organisational relationships from trivial purchases to major national infrastructure projects. Any large organisation (whether private or public) must unavoidably invest significant resources in developing and concluding contracts, as the contracts it enters define its legal relations with organisations and individuals with which it interacts. As one author has noted, contracts are an integral part of any business enterprise and it is “difficult to overstate their importance to the business world” (Khouri and Yamouni, 2007, p16). In addition, drafting contracts represents a major economic activity for the legal industry.

Ambiguity is an unresolved problem in natural language processing. It is also an important

problem affecting the drafting and operation of contracts. The costs of ambiguity in a contract can be enormous. The presence of ambiguity increases the risk of litigation, project failure and loss of commercial relationships. Instances of litigated ambiguity are used in drafting manuals as instructional tales for the unwary drafter (Aitken and Butt, 2004, p37).

Our future research goal is to investigate approaches that will help identify such ambiguity, however, a precursor for addressing the problem of ambiguity is accurate labelling or classification of the constituent parts of contract documents.¹ In this paper we report work on classifying ‘lines’ in contracts into 32 classes. We compare three different classification approaches in order to determine which approach may provide the best performance with respect to this task: supervised machine learning; a hand-crafted rule based tagger; or combining hand-crafted rules based tagging with machine learning.

2 Background

Natural language processing involves applying a pipeline of transformations to text. A typical first step is to segment the text into its sentences, after which further processing is applied (Jurafsky and Martin, 2009, p 69). This works well in usual English prose, but is not necessarily well suited to contracts because of the heavy use of layout to embody structure and semantics. In a contract, a sentence may occur within a single paragraph of text, or may be spread over several line breaks due to the use of sub-paragraphing as an aid to comprehension. Such sub-paragraphs may embody disjunctive or conjunctive conditions associated with

¹This task can also be thought of as a document segmentation task, although we here consider it as a classification exercise.

a rule. A ‘clause’² is the typical structural unit which embodies a legal rule or set of rules. Sometimes clauses are organised into sub-clauses and clauses and sub-clauses may be numbered in hierarchical fashion (e.g. 1, 1.1, (a), (i), (A))(Aitken and Butt, 2004, pp 23, 27). Line breaks often occur immediately before data such as headings, clauses, sub-clauses, party names, dates or execution blocks. Being able to label a line as to its character, key content or function thus enriches available information for later NLP. The line is thus a valuable starting point for applying NLP.

3 Related Work

3.1 Classification and Segmentation of Contracts and Legislative Documents

There would appear to be little published work on the application of natural language processing (NLP) or machine learning specifically for the classification of text in legal contracts. An exception to this is (Indukuri and Krishna, 2010) who apply a support vector machine over feature vectors extracted from contract clauses. These feature vectors are comprised of n-grams up to size $n = 4$. They classify contract sentences as clauses or non-clauses and then sub-classify clauses as concerned with payment terms or not. Indukuri and Krishna note the lack of published work on classification of sentences in contract documents for workflow and monitoring purposes.

While also limited, there has been some work on segmentation or classification of legislative corpora using machine learning. (Mencía, 2009) reports 100% precision and recall in machine learning on classification of legislation into articles, sections and parts (although to a task for which he is able to craft an equally accurate regular expression based segmenter). (Bacci et al., 2009) automatically classify plain text legislative documents for the purpose of XML mark up. (Francesconi, 2006), on whose work Bacci et al. build, describes four separate modules for dealing with plain text legacy content which they are seeking to convert to XML tagged text. (Hasan et al., 2008) carry out segmentation of Spanish legislative bulletins using the table of contents as an aid to segmentation. This work, in the legislative field, provides a parallel application domain for the work undertaken here in respect of contracts, although it is

²The term ‘clause’ here refers to a legal rule or rules appearing in a contract, rather than the linguistic entity.

generally true that legislation will tend to conform to far tighter stylistic rules, because of the central control of legislative drafting.

3.2 Combining Machine Learning and Hand Coded Approaches

A number of researchers have combined hand-coded approaches and machine learning to produce better overall results.

(Kipp, 2006) augments handcrafted rules with machine learnt rules for gesture generation.

(McCreath and Kay, 2003) used such a combined approach to improve performance on categorizing emails.

(Park and Zhang, 2003) address the problem of text chunking for the Korean language, combining hand-crafted rules with a memory based machine learning method. A rule based method is used to determine whether an instance may constitute an exceptional case, with machine learning used to correct classifications assigned by the rule based method. They report up to a 2.83% improvement in F-scores over memory-based learning alone.

(Rochery et al., 2002) report the use of hand-crafted rules to enhance the accuracy of machine learning utilising boosting. They apply the method to a multi-labelling classification problem where they are seeking to classify spoken utterances. They attribute increased accuracy to the fact that the rules they develop are not represented in their data at all or do not appear sufficiently to have statistical impact during the training process. They note that the combination has the greatest effect when there are less than 100 examples, with the benefit of the combination decreasing as the number of training examples increases. The method therefore has direct applicability to reducing the data requirement for machine learning (a costly aspect of supervised machine learning).

(Takahashi et al., 2005) utilise a combination of SVM based learning and hand-crafted rules for the task of classifying occupations from social surveys. They effect a combination by using rule derived labels as features for the machine learner. Again the work occurs in a multi-class labelling context (with around 200 occupation codes). They find that SVM approaches are superior to rule-based approaches alone but that combining rule-based and SVM learning produces the best results. They also examine the effect of the number of training examples, noting that the differential ben-

efits of combining methods reduces with increasing data.

In our study we use class labels assigned by a hand-coded tagger as additional features for machine learning to obtain an increase in accuracy in the domain of text classification.

The work reported above, and our own paper, come from diverse domains and illustrate the application of combined rule/ML approaches using a variety of machine learning algorithms.

This suggests the wide utility of using hand-coded rules to improve accuracy of machine learning. Symetrically we may say that machine learning may potentially be used to improve the accuracy of hand-coded rules. Here we essentially follow a hybrid model of development that examines each method separately as well as in combination.

4 Data

4.1 Australian Contract Corpus

The data which forms the basis of this study is a corpus of 256 contracts ('the Australian Contract Corpus') which has been compiled from the web using the Google Australia web search: <http://www.google.com.au>.³ The corpus is constituted of 1043364 tokens (words and punctuation), 42910 sentences, and a vocabulary of 14217. The lexical diversity of this corpus is 73.39.⁴ As far as we are aware there is no published work based on a corpus of Australian contract language, however space does not permit a fuller discussion of the design and analysis of the corpus.

4.2 Data Representation

For the purpose of this study we randomly selected 30 contracts from the Australian Contract Corpus to produce data sets for application of machine learning. The 30 contracts were divided into three

sets: a training set (Set A)⁵, a second set (Set B)⁶ (primarily used as a test set, although sometimes combined with Set A to form a larger training set) and an additional test set (Set C)⁷, each of 10 contracts. Set C was added to increase both available data and the number of contracts on which testing could be carried out.

To simplify processing all contracts were converted into text files as a preprocessing step.⁸ Also we removed material that typically does not appear in contracts (primarily guideline notes for drafters).

For purposes of our study each 'line' or 'paragraph' in a contract constituted a data point. Sets A, B and C provided 1825, 2157 and 2231 lines/data points, respectively.

To classify our data for machine learning we first employed our hand-coded tagger to produce a labelled data set constituted of a line and primary and secondary labels according to our classification system (see below). The classifications applied by the hand-coded tagger were then manually corrected to remove any labelling errors. Both for the reason that hand tagging data from scratch is more error prone than correcting previously machine tagged data, and because of the high labour in manual tagging, machine assisted tagging is a standard method used in corpus creation (Biber et al., 1998, p 262), ((ed) Christopher S. Butler, 1992, p 131).

4.3 Feature Selection

Features were then extracted from our labelled lines using a hand-coded feature extractor to extract up to 40 features to be used for machine learning, with the primary class label earlier applied to each line serving as the target for machine learning.⁹ Features for machine learning were selected based on assumed relevance to the intended classification.¹⁰ While a bag of word representa-

³The search phrase we employed was: "clause party agreement", limiting the search to "pages from Australia" and the filetype to ".doc". The collection of the corpus was undertaken in the period 6 - 24 December 2009. Each document was visually inspected by one of the authors to verify that it was an Australian contract and documents were added to the corpus in order of their appearance in Google search results. We have made available the list of URLs of the documents that make up the corpus at: <http://cs.anu.edu.au/people/Michael.Curtotti/>. We also make code and data referred to in this paper available at the same location.

⁴Lexical diversity refers to the ratio of the total number of tokens in the corpus to the total number of types of tokens i.e. tokens/vocabulary. The statistics and information provided here were extracted using the NLTK. (Bird et al., 2009).

⁵Contracts numbered in our corpus: 55, 74, 77, 91, 94, 144, 174, 185, 208 and 213

⁶Contracts 11, 12, 164, 193, 196, 199, 249, 59, 64, 9

⁷Contracts 200, 254, 175, 180, 120, 102, 207, 127, 75 and 79

⁸Notably, this step does have the downside of discarding valuable style and layout information found in the word document format.

⁹In some instances the assignment of class is not entirely disjunctive and implies a priority ordering of potential classifications. For instance a clause may contain address details, and be classified as such rather than as a clause. The feature extractor removed secondary labels.

¹⁰The full list of features is as follows: 1. relative position of a line within the contract; 2. line end punctuation; 3. num-

tion with automatic feature extraction was trialed, this was found to result in lower accuracy than a manually selected (but small) number of features. Traditionally, ‘feature engineering’ or ‘selection’ in text classification has focussed on ‘bag of words’ or proximate features such as n-grams or noun phrases (Scott and Matwin, 1999). Our selection of features might be regarded as ‘hybrid’ in that they implicitly encode domain knowledge and ‘rules’. For example, we expect a definition line to contain the widely used word ‘means’ as in ‘x means y’. We would expect clause headings to display a different pattern of parts of speech occurrence, as compared to clauses themselves. We also consider features such as the relationship between the particular line/datapoint and its neighbouring lines (e.g. is the preceding line long or short), or the document as a whole (e.g. relative placement of the line within the document).

Features used in the related work carried out by others on contracts and legislation include: n-grams (Indukuri), bag of word together with capitalisation and character patterns (Mencia), the use of sequential word and token data as a HMM to predict state where state is the classification of the portion of text (Bacci/Francesconi) and the use of word terms identifying titles and lists of indexed terms (Hasan).

4.4 Classes

We adopted 32 classes for our classification scheme. These classes represented significant

being at the beginning of the line; 4. line length in characters; 5. line length in tokens; 6. number of nouns; 7. number of adjectives; 8. number of verbs; 9. number of prepositions; 10. number of coordinating conjunctions and cardinal numbers; 11. number of modal verbs; 12. number of personal pronouns, possessives; 13. number of adjectives; 14. whether must occurs; 15. whether may occurs; 16. whether shall occurs; 17. whether means occurs; 18. the position of the word means in the line; 19. whether the word ‘include’ occurs; 20. the position of the word ‘includes’ in the line; 21. whether the phrase ‘has the same meaning’ occurs; 22. whether the line begins with a capital letter followed by a stop or a space; 23. do the letters ‘abn’ or ‘acn’ appear; 24. does the word address appear near the beginning of the line; 25. does the word ‘contact’ appear near the beginning of the line; 26. does the word ‘email’ appear near the beginning of the line; 27. does the word ‘fax’ appear near the beginning of the line; 28. does the word ‘note’ appear at the beginning of the line; 29. does the word ‘phone’ appear at the beginning of the line; 30. do the terms ‘web’ or ‘www’ appear near the beginning of the line; 31-36. token lengths of 3 lines before and after the data line; 37. the tag applied by the hand coded tagger to the data line; 38. whether the contract from which the line comes has clause headings; 39. whether the contract from which the line comes has clause sub-headings; 40. whether the contract has a schedule.

structural elements (such as clause headings or content lines) or sometimes key contract meta data (for instance the parties, the date on which an agreement is made, email addresses, ABNs etc).¹¹ Of particular interest to us, was an ability to accurately identify clausematter (the primary location of legal rules of a contract) and clause headings (which in some contracts effectively mark boundaries between major rule sets).

5 Experimental Evaluation

Analysis of our data sets was carried out using two major methods: analysis using the training Set A measuring performance on test Set B (i.e. testing for accuracy of classification of lines in Set B); and analysis using training Set A where each contract in test Sets B and C was used individually as a test set (testing for accuracy of classification of lines of each individual contract). The former testing approach is more widely used but the latter is more consistent with our intended end application: that is, developing an ability to accurately classify lines in a previously unseen contract as an individual document.

Set A consisted of 1825 lines of data, Set B consisted of 2157 lines of data and Set C consisted of 2231 lines of data.

5.1 Tools

The following tools were used to carry out machine learning tasks reported in this paper:

1. the *python* programming language was used to develop a hand coded line tagger, a feature extractor and code for evaluation of the performance of the hand coded tagger;
2. the Natural Language Toolkit (NLTK)(Bird et al., 2009) was used in corpus development and in the hand coded tagger for application of parts of speech tags; and
3. the Weka data mining software(Hall et al., 2009) was used to carry out machine learning.

¹¹The following class labels were applied: CLAUSE-MATTER, CLAUSEHEAD, DEF, CONTENTSHEAD, BLANK, DROSS, AND, BETWEEN, DATEMADELINE, PARTIESHEAD, PARTYLINE, RECITALHEAD, RECITALLINE, HEAD, CONTENTLINE, PRELIM, OPERATIVEPROVISIONLINE, NUMBEREDLINE, SCHEDULEHEAD, SCHEDULEITEM, EXECUTIONBLOCK, NOTE, ABNLINE, ADDRESSLINE, EMAILLINE, WEBSITE, FAXLINE, PH-LINE, REF-LINE, SCH-OTHERMATTER, CONTACTOFFICER, TITLELINE.

5.2 Hand-Coded Tagger

Our hand-coded tagger follows a pipeline methodology. A contract is separated into a list of lines. In the first phase, each line is passed through a series of “if-then-else” routines (essentially utilising regular expressions) to assign primary and secondary labels to the line. This produces a contract with enhanced information content, in the form of a tuple of labels and line text. This output is then fed through a series of methods, progressively bootstrapping improvements to labelling, based on the information available when the method is applied. For instance content lines will be identified, these are then used in later methods to seek to identify clause headings. In addition, the tagger extracts contract meta data such as: a list of definitions, a list of clause headings, where the beginning of operative provisions occurs, where the schedule begins. Such features may or may not occur in a contract. In addition the tagger also provided a method for hand correcting tagging which was used to create our data set for supervised machine learning purposes.

Development of the tagger relied heavily on domain knowledge to identify logical tests for label application. For instance the occurrence of numbering at the beginning a line was used to identify clause headings, or the presence of a common text pattern which occur in definitions i.e. (the pattern *[“word[1-3]] means [...]”* was used as a criterion for identifying a line as a definition).

	$\bar{x}$	s
On Set B Contracts	86.27%	NA
Contract x Contract	82.84%	10.85%

Table 1: Accuracy of Line Tagger.

The first row of Table 1 reports accuracy of the hand coded tagger when applied to Set B as the test set (the % accuracy of tagging of individual lines in Set B is reported). The second row of Table 1 reports average accuracy with which the lines in each of 20 contracts in Sets B and C are tagged (standard deviation is also reported). The high variance of the results at a contract by contract level suggests a need for future work to identify if issues such as contract ‘type’ may account for the differences (the lowest level of accuracy was around 50% and the highest around 95%).

Table 2 provides measures of precision, recall and F-measure attained by the hand coded tag-

	Precision	Recall	F-Measure
CL.HEAD	0.98	0.83	0.89
CL.MATTER	0.91	0.91	0.91
DEF	1.00	0.66	0.80

Table 2: Performance results for Hand-Coded Line Tagger

ger on Set B for key data items: clause headings, clauses themselves and definitions.¹² These results on key measures are sufficient for our initial purposes and will allow us to focus subsequent work on addressing ambiguity in the substantive rules found in a contract.

5.3 Supervised Machine Learning

Column 1 of Table 3 reports the accuracy of various supervised machine learning algorithms with Set A as training set and Set B as test set. Of the algorithms applied, Random Forest (implemented using 100 trees and 30 random features) performed best. No ML method performed better than the hand coded tagger tested on the same test set (see row 1 of Table 1).

5.4 Machine Learning vs Hand-coded methods

Given our ultimate focus on developing applied tools, we are primarily concerned to identify the best methods to maximise the accuracy of our classification task. Whether this proves to be machine learning methods or hand-coded methods or a combination of the two is immaterial from this viewpoint (other than in regard of development costs).

Also it is useful to consider whether methods and insights from one method may assist in the other. For instance, hand-coded rules that successfully labelled lines directly suggested ‘features’ to be extracted for machine learning purposes. Conversely the identification of ‘features’ for machine learning can be regarded as a programming abstraction where the algorithm is ‘under the hood’ but human intervention is required in the form of feature selection for ‘learning’ to occur (implicitly encoding of rules in the feature set). In deriving our results we used a confusion matrix as a tool to assist in identifying areas where the hand-

¹²Precision is $\frac{TP}{TP+FP}$, Recall is $\frac{TP}{TP+FN}$, and F-Measure is $\frac{2PR}{P+R}$, where TP is true positives, FP is false positives, FN is false negatives, P is precision, and R is recall.

coded tagger required improvement and as a common baseline for performance. In certain forms of machine learning, the resulting machine learning model can be represented as code in “if, then, else” form (as in the output of a decision list or decision tree learner). We may thus conceive of machine learning and hand coding of rules as different ‘programming paradigms’ - a perspective that encourages us to apply insights across the boundary between them.

Further, we wished to explore the relative accuracy of machine learning vs hand coding for this particular task. Such an exploration is valuable from a cost benefit point of view. Supervised machine learning (where we have focused our attention) requires the production of human annotated data (which is costly to produce) as well as feature testing and refinement. Also machine learning may be computationally intensive. Production of an accurate hand-coded tagger requires expert knowledge and considerable testing and code refinement in order to achieve a high level of accuracy (which again is resource intensive). We did not undertake quantitative data collection in respect of time employed in the two methods, particularly as we followed an iterative development model that switched between rule based and machine learning focussed work and as we were searching for optimal outcomes. Nonetheless we anecdotally report that each method presented significant demands in terms of development time.

In the results reported above we note that no machine learning approach alone exceeded the performance of the hand coded tagger when considering performance on test Set B as a whole. This is the result also found at contract by contract level (see below Table 4).

5.5 Leveraging Hand-Coded Labels for Machine Learning

As noted above we also wished to explore whether utilising the output from the hand coded tagger as input features for machine learning might improve the accuracy of machine learning alone. A converse question that could be framed is whether the rule based tagger’s performance could be improved by combination with machine learning.

We carried out two trials that bear on these questions:

1. machine learning alone and machine learning enhanced with rule based input on an entire

test set (Set B); and

2. machine learning alone, rule alone and machine learning enhanced with rule based input at contract by contract level.

ML Algorithm	ML Alone	ML + Rule
Naive Bayes	82.38%	84.75%
SMO (SVM)	82.80%	85.40%
Cl. via Regression	83.91%	87.07%
Decision Tree	82.01%	87.02%
Bagging	83.54%	87.76%
Majority Vote	84.42%	87.11%
Random Forest	85.12%	86.69%

Table 3: Comparative Performance ML alone and ML + Rule on train and test set

Column 2 of Table 3 shows the improvement in performance of various machine learning methods when tags applied by the hand coded tagger were provided as input features. All methods show improvement with addition of the rule derived feature on Set B as a whole.¹³

We report below a comparison of accuracy of machine learning, rule based tagging and a combination of both in respect of Decision Trees. Decision trees offer the advantage that the output is easily interpretable as a rule set for tagging.

Method	$\bar{x}$	s
ML Alone	79.12%	10.71%
Rule Alone	82.84%	10.85%
ML + Rule	83.48%	9.83%

Table 4: Tagging Accuracy Mean and Standard Deviation - Contract by Contract

Although the results reported in Table 4 are different (as different contracts are included in the tested contracts)¹⁴ average accuracy continued to show a differential between machine learning alone, and machine learning leveraged with hand rules, as described above. On this combination ML + Rule also outperformed the rule based tagger alone, but by a very narrow margin. In 17 of 20 cases ML + Rule improved or did not worsen performance of ML alone. In 15 of 20 cases, ML +

¹³The training set for the results reported in Table 3 was Set A. The test set was Set B.

¹⁴The training set was Set A. Sets B and C provided 20 individual contracts (effectively 20 test sets) for testing.

rule was no worse or outperformed the hand coded tagger alone. Notably variation is quite high as compared to the mean. Testing for statistical significance we can reject a null hypothesis that the mean accuracy for ML + Rule is the same as for ML alone. We cannot reject a null hypothesis that ML + Rule is the same as Rule alone.¹⁵ We note also that this investigation was only carried out in respect of the decision tree learner.

5.5.1 Is it more effective to increase data or use rule based features?

To further explore the question of how much effort might be saved by combining rule based approaches with machine learning alone we tested the effect of adding random noise to the feature set provided by the hand coded tagger. We found that we had to randomly flip over 45% of these tags before they ceased to impact positively on classification accuracy. This suggests that even poorly prepared hand coded rules can improve performance.¹⁶

We also tested the effect of incrementally increasing the amount of available training data to explore any decrease in differential in performance. Testing using Set C as the test set and taking training data from Sets A and B, we progressively increased training data by increments of 500 instances of data. Differential performance decreases progressively, starting around 5% and reducing to 2.5% as more training data is made available.¹⁷ (See Table 5.) Extrapolating from these figures we would expect at a minimum that training data would have to be doubled to remove differential accuracy. Relatively little effort in developing rule based features may substitute for considerable work in creating additional supervised data.

¹⁵We gratefully acknowledge the assistance of Bob Forrester of the ANU Statistical Consulting Unit in deriving this result, undertaken utilising a generalised linear model to undertake model fitting adjusting for different file (contract) types. Predictions of estimated mean proportions were: ML Alone 0.7760 (s.e. 0.007181), Rule alone 0.8106 (s.e. 0.006775), ML+ Rule 0.8181 (s.e. 0.006677).

¹⁶Using a Decision Tree, ML + Rule accuracy on test Set B with training Set A, was 87.0% with zero noise, 86.9% with 5% noise, 84.7% with 10% noise, 85.2% with 20% noise, 84.4% with 30% noise, 83.9% with 40% noise, 82.8% with 45% noise, 82.4% with 50% noise. Decision Tree accuracy on ML alone on this training and test set was 82.0%.

¹⁷The data has a negative correlation of -0.70

Data	0.5	1K	1.5	2K	2.5	3K	3.5	4K
Diff.	5.1	3.4	5	4.2	3.3	3.1	3.9	2.6

Table 5: Difference in % accuracy of ML alone and ML with rule based feature with increasing training data.

5.5.2 Decision tree output

An examination of the decision tree produced after adding the output of the hand-coded tagger to machine learning is of interest.

```

LTlabel = #CLAUSEMATTER#
| tknLngth <= 4
|| nounNum <= 1: #CLAUSEMATTER# (6.0/2.0)
|| nounNum > 1
||| linePos. <= 0.383929: #PRELIM# (2.0)
||| linePos. > 0.383929: #CLAUSEHEAD# (2.0)
| tknLngth > 4: #CLAUSEMATTER# (508.0/7.0)
: :
: :
LTlabel = #PH-LINE#
| prepNum <= 1: #PH-LINE# (5.0)
| prepNum > 1: #CLAUSEMATTER# (2.0)
LTlabel = #EMAILLINE#: #EMAILLINE# (3.0)
LTlabel = #ABNLINE#
| preLine2 <= 0: #TITLELINE# (2.0/1.0)
| preLine2 > 0: #ABNLINE# (3.0)

```

Figure 1: Part of the decision tree learnt when the labels from the hand-coded are provided as an attribute.

Using only the machine learning feature set, the size of the pruned tree is 199 with 108 leaves. After addition of labels provided by our hand-coded tagger the size of the decision tree is reduced to 105 with 69 leaves. Figure 1 shows part of this decision tree and as illustrated, the decision tree generally begins with the label assigned by the hand-coded tagger 'LTlabel' and amends the label (where necessary) utilizing other features available to the machine learner. Some labels (e.g. EMAILLINE) were simply adopted without change. In the case of the CLAUSEMATTER label, the machine learner used the number of tokens in the line ('tknLngth'), number of nouns in the line ('nounNum') and the relative position of the line in the contract ('linePos.') to re-assign the class of the line. The significance of the use of labels assigned by the hand-coded tagger is that the decision-tree identifies that attribute as providing the most significant information gain (that is reduction in entropy) and therefore partitions on that feature (Callan, 2003, p245). This remains true in this example irrespective of the class to which

the data is being assigned.¹⁸ It is obvious that the best feature for assigning the correct label for an item of data is the correct or probably correct label (if one can obtain it without undue endeavour). From a practical viewpoint some effort directing to hand-crafting code to extract appropriate class labels may thus assist machine learning.

6 Conclusions and Future Work

In terms of the overall performance of the methods described in this paper the closest point of comparison is the work reported by Indukuri and Krishna (Indukuri and Krishna, 2010). While comparisons using different methods, software and data are of dubious validity, we may note their broad similarity. With a training set of 10 contracts and a test set of 20 contracts, the contract-by-contract average F-measure for the clausematter class was 0.8632. Also in 14 contracts of the 20 tested the F-measure for the clausematter class was above 0.85. With a training set of 10 contracts (Set A) and a test set of 10 contracts (Set B) (as a single test set) the F-measure for the clausematter class was 0.921.¹⁹ Using 4-grams as features, Indukuri and Krishna report an accuracy of 83.48% on the task of classification clauses from non-clauses, applied to 73 sentences to be so classified. Whereas the task in that case was a binary classification task over one contract, we report results on multi-class classification over test sets of up to 20 contracts.²⁰

The work we have undertaken and related work reviewed in this paper (Section 3) suggests the potentially broad utility of combining rule based and machine learning methods. One potential hybrid classification method in its simplest formulation may look something like this:

1. determine the required or anticipated level of accuracy for the intended application;
2. develop a simple code rule set for the classification task (i.e. with minimal expenditure of resources);
3. if classification accuracy is sufficient machine learning would not be required, if not

¹⁸The Weka decision tree implementation, “J48”, is based on the Quinlan’s C4.5(Quinlan, 1986).

¹⁹The results reported here are derived using a random forest algorithm and using the hand coded tag as an input feature for classification.

²⁰Coincidentally the 83.48% figure is the same as our average result over all classes for 20 contracts.

proceed as usual in development of machine learning using the output of hand coded rules as an input to machine learning.

Such a development model does not avoid the development of a supervised training set, as such a set is required both to assess the accuracy of the rule based tagger as well as the machine learner (should development proceed to that stage). However it may reduce the amount of data required to attain a desired level of accuracy. The method could of course be applied iteratively, if the accuracy level is insufficient after a first iteration.

Future work to be explored includes the effective “typing” of contracts (particularly in light of the high variance of performance on individual contracts). For instance, some contracts use clause headings, others do not. Some use schedules, others do not, etc. Such type information may further assist in the classification task (the accuracy of which varies considerably depending on the contract to which it is applied). In typical preprocessing, documents in corpora are converted to plain text. This results in the loss of layout and hierarchical information found in word documents. The preservation of such information for use in classification tasks may improve accuracy.

The preliminary work here is also a precursor to work on identification of ambiguity and the development of practical tools to assist in contract drafting. In separate work, we intend also to describe and analyse the contract corpus that is referred to here as the basis of this work.

References

- J. K. Aitken and P. Butt. 2004. *Piesse The Elements of Drafting*. Lawbook Co, 10th edition.
- L. Bacci, P. Spinosa, C. Marchetti, R. Battistoni, I. Florence, I. Senate, and I. Rome. 2009. Automatic mark-up of legislative documents and its application to parallel text generation. In *Proc. of LOAIT Workshop*, pages 45–54.
- D. Biber, S. Conrad, and R. Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.
- Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O’Reilly Media Inc.
- Rob Callan. 2003. *Artificial Intelligence*. Palgrave Macmillan.

- (ed) Christopher S. Butler. 1992. *Computers and written texts*. Blackwell.
- E. Francesconi. 2006. The Norme in Rete Project: Standards and Tools for Italian Legislation. *International Journal Legal Information*, 34:358.
- Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software. *SIGKDD Explorations*, 11(1).
- I. Hasan, J. Parapar, and R. Blanco. 2008. Segmentation of legislative documents using a domain-specific lexicon. In *Proceedings of the 19th International Conference on Database and Expert Systems Application*, pages 665–669.
- Kishore Varma Indukuri and P. Radha Krishna. 2010. Mining e-contract documents to classify clauses. In *COMPUTE '10: Proceedings of the Third Annual ACM Bangalore Conference*, pages 1–5, New York, NY, USA. ACM.
- Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, 2nd edition.
- D. Khoury and Y. S. Yamouni. 2007. *Understanding Contract Law*. Lexis Nexis Butterworths, 7th edition.
- M. Kipp. 2006. Creativity meets automation: Combining nonverbal action authoring with rules and machine learning. In *Intelligent Virtual Agents*, pages 230–242. Springer.
- Eric McCreath and Judy Kay. 2003. Iems: Helping Users Manage Email. *User Modeling 2003*, pages 146–146.
- Eneldo Loza Mencía. 2009. Segmentation of legal documents. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 88–97. ACM.
- Seong-Bae Park and Byoung-Tak Zhang. 2003. Text chunking by combining hand-crafted rules and memory-based learning. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 497–504.
- J.R. Quinlan. 1986. Induction of decision trees. *Machine Learning*, 1:81–106.
- M. Rochery, R. Schapire, M. Rahim, N. Gupta, G. Ricciardi, S. Bangalore, H. Alshawi, and S. Douglas. 2002. Combining prior knowledge and boosting for call classification in spoken language dialogue. In *Proceedings of the IEEE International Conference of Acoustics and Speech and Signal Processing*.
- S. Scott and S. Matwin. 1999. Feature Engineering for Text Classification. In *Machine learning: proceedings of the Sixteenth International Conference (ICML'99), Bled, Slovenia, June 27-30, 1999*, page 379. Morgan Kaufmann Pub.
- K. Takahashi, H. Takamura, and M. Okumura. 2005. Automatic occupation coding with combination of machine learning and hand-crafted rules. *Advances in Knowledge Discovery and Data Mining*, pages 269–279.

APPENDIX

1008C

Citations to Curtotti

Source document:

Curtotti, M., & McCreath, E. (2010, December). Corpus based classification of text in Australian contracts. In *Proceedings of the Australasian language technology association workshop*, pp. 18-26.

Citations: 2013 – 2017 (in chronological order)

Gao, X., & Singh, M. P. (2013, March). Mining contracts for business events and temporal constraints in service engagements. *IEEE Transactions on Services Computing*, 7(3), 427-439.

Gabbard, J., Sukkarieh, J. Z., & Silva, F. (2015, June). Writing and reviewing contracts: don't you wish to save time, effort, and money?. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law* (pp. 229-230).

Bateman et al. (2015, October). Natural Language Processing for Extracting Conveyance Graphs, United States Patent Application Publication, no. US 2015/0286630 A1.

Aires, J. P., & Meneguzzi, F. (2017, May). Norm conflict identification using deep learning. In *International Conference on Autonomous Agents and Multiagent Systems* (pp. 194-207). Cham: Springer International Publishing.

Cardellino, C., Teruel, M., Alemany, L. A., & Villata, S. (2017, June). A low-cost, high-coverage legal named entity recognizer, classifier and linker. In *Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law* (pp. 9-18).

Chalkidis, I., Androutsopoulos, I., & Michos, A. (2017, June). Extracting contract elements. In *Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law* (pp. 19-28).

Engelfriet, A. (2017, September). Creating an Artificial Intelligence for NDA Evaluation. Available at SSRN: <https://ssrn.com/abstract=3039353> or <http://dx.doi.org/10.2139/ssrn.3039353>

APPENDIX

1011A

← → ↺ web.archive.org/web/20230202142404/https://aclanthology.org/2022.findings-emnlp.447.pdf ☆

INTERNET ARCHIVE <https://aclanthology.org/2022.findings-emnlp.447.pdf> Go JAN FEB 02 MAR 2022 2023 2024 About this capture

2 captures 2 Feb 2023 - 26 Mar 2025

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. Israelson, Clayton T. Morrison,
Derek E. Bambauer, Mihai Surdeanu

The University of Arizona, Tucson, Arizona, USA

{alicekwak, jisraelson, 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

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-

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

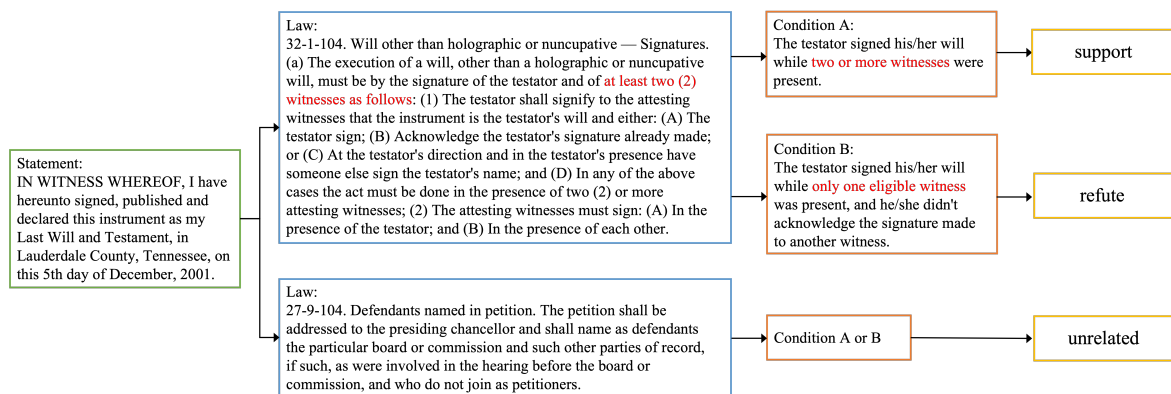


Figure 1: A demonstration of the annotation procedure. Given a law and condition, each statement is evaluated for validity. *Support*: it is supported by the law and condition provided; *Refute*: refuted by the law and condition. *Unrelated*: not relevant to the law. When a statement and a law are unrelated 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 OCR³, and copying statements into the dataset;
- (2) for each statement, identifying the Tennessee

²A 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.

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

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

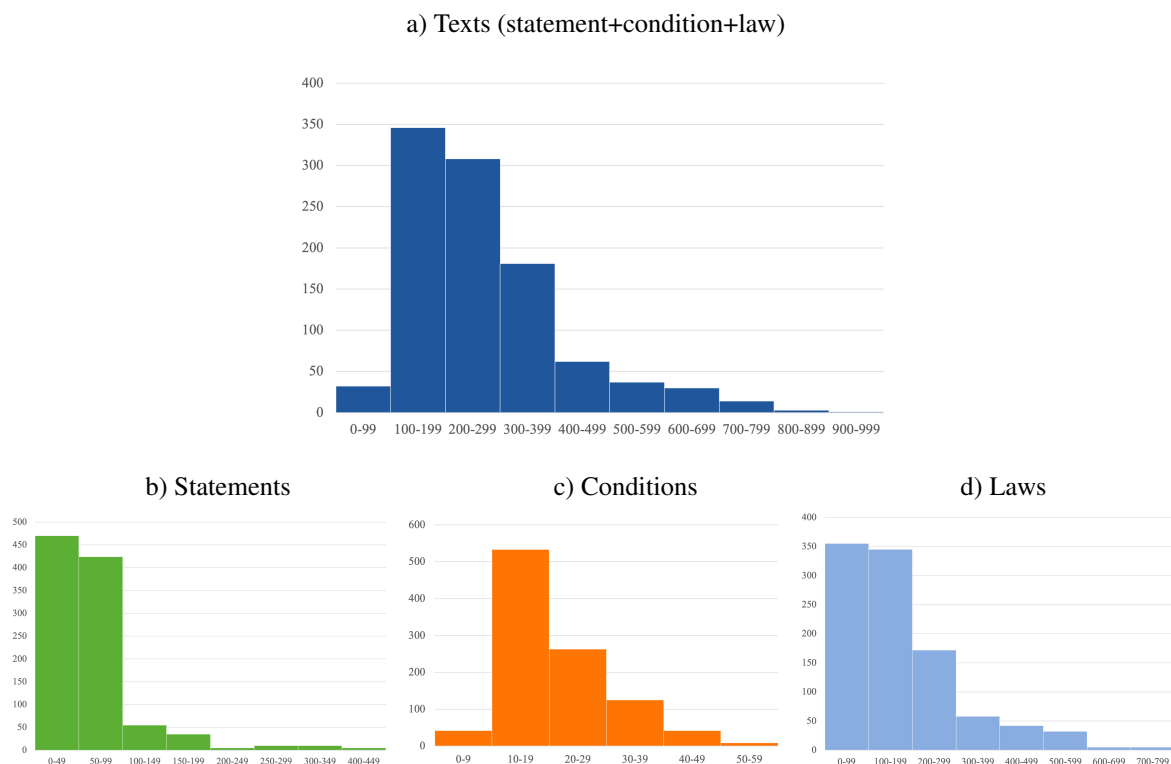


Figure 2: Histograms demonstrating the length distribution of a) full texts (i.e., statement+condition+law), b) statements, c) conditions, and d) laws in terms of token counts. Token counts are plotted on the X-axis; the Y-axis indicates the number of texts/statements/conditions/laws for each bin.

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 distilbert-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 page⁴) were chosen: all-mpnet-base-v2⁵, multi-qa-mpnet-base-

⁴<https://www.sbert.net>

⁵<https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

dot-v1⁶, 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

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>

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

⁹For 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.

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 perfor-

mance. 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

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 predictions¹⁰. Laws have more impact on performance than conditions, since deterioration in results without laws was more significant than from models trained without condi-

tions. 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

¹⁰A 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.

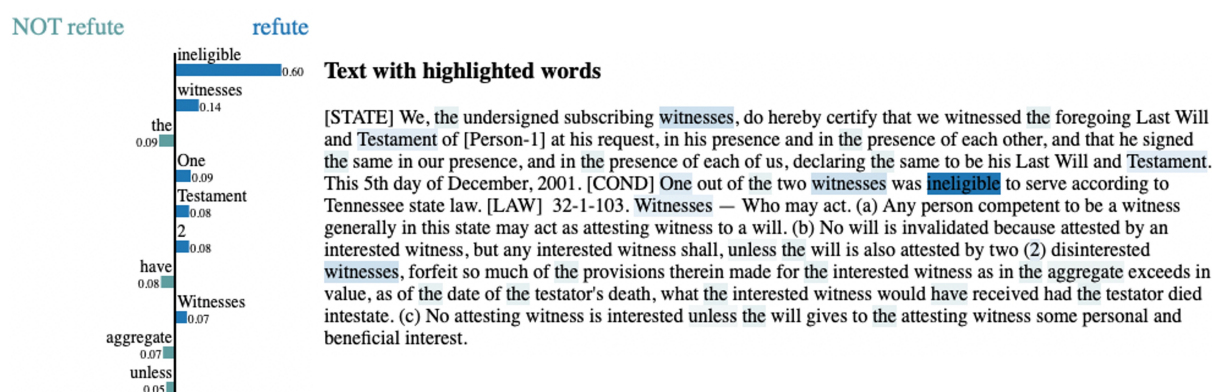


Figure 3: Examples of LIME explanations showing a case where the model (roberta-large-mnli) makes correct predictions based on features sensible to humans.

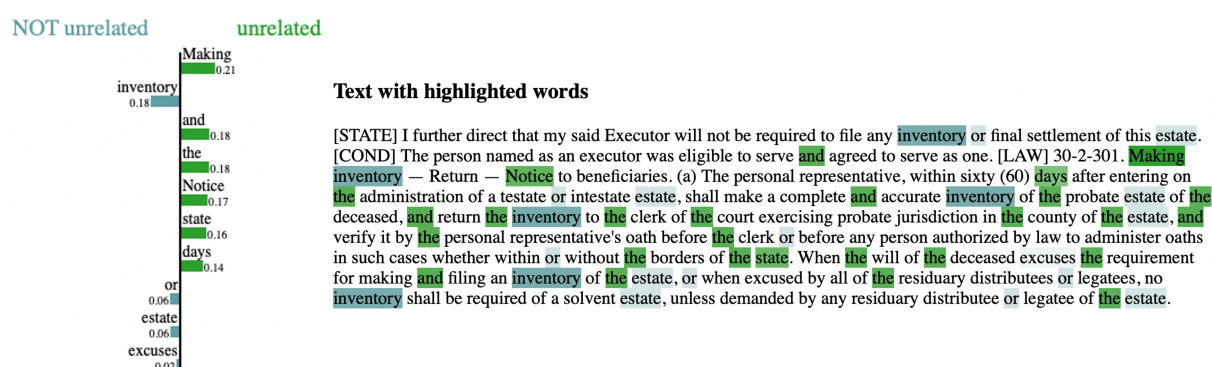


Figure 4: An example of a LIME explanation showing a case where the model (roberta-large-mnli) makes an incorrect prediction based on the tokens such as *Making*, *and*, and *the* which bear little semantic relevancy to the gist of the text.

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.

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 Textual 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 Sunawal, 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. [IFly-Legal: 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

1011B

Validity Assessment of Legal Will Statements as Natural Language Inference

Alice Saebom Kwak, Jacob O. Israelson, Clayton T. Morrison,
Derek E. Bambauer, Mihai Surdeanu

The University of Arizona, Tucson, Arizona, USA

{alicekwak, jisraelson, 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

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-

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

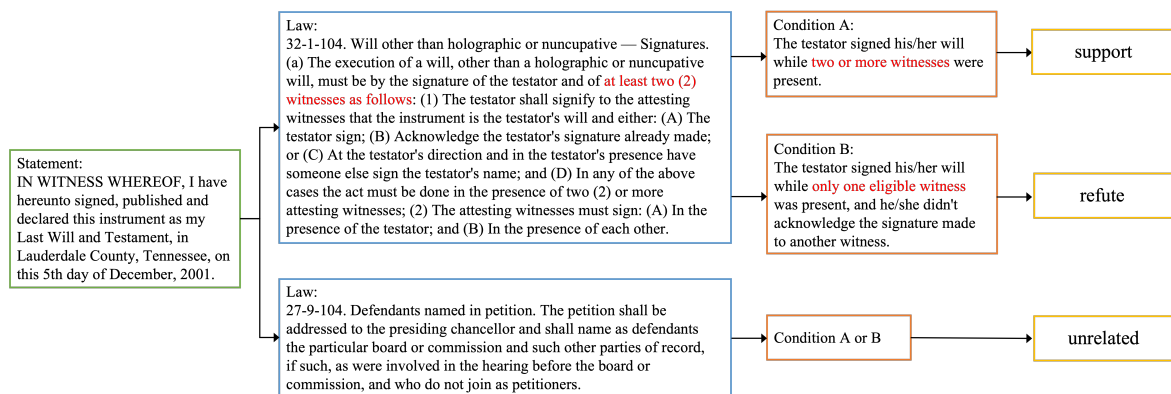


Figure 1: A demonstration of the annotation procedure. Given a law and condition, each statement is evaluated for validity. *Support*: it is supported by the law and condition provided; *Refute*: refuted by the law and condition. *Unrelated*: not relevant to the law. When a statement and a law are unrelated 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 OCR³, and copying statements into the dataset;

(2) for each statement, identifying the Tennessee

²A 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.

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

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

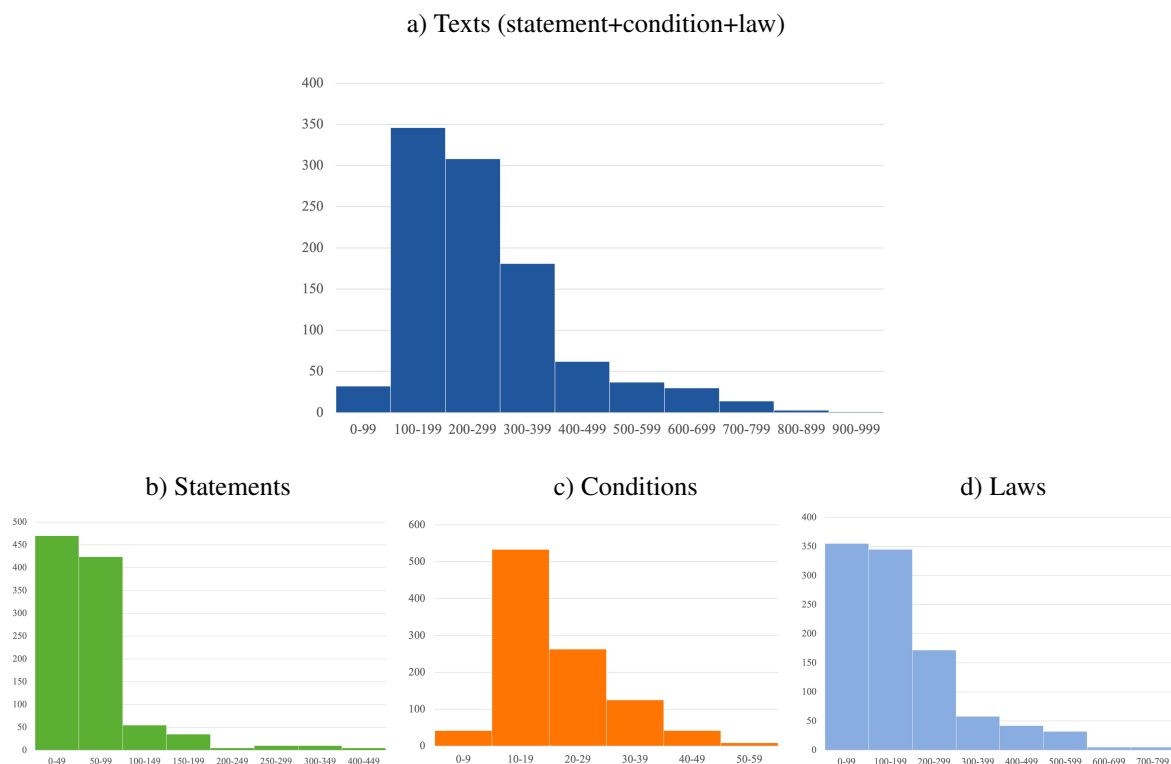


Figure 2: Histograms demonstrating the length distribution of a) full texts (i.e., statement+condition+law), b) statements, c) conditions, and d) laws in terms of token counts. Token counts are plotted on the X-axis; the Y-axis indicates the number of texts/statements/conditions/laws for each bin.

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 distilbert-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 page⁴) were chosen: all-mpnet-base-v2⁵, multi-qa-mpnet-base-

⁴<https://www.sbert.net>

⁵<https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

dot-v1⁶, 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

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>

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

⁹For 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.

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 perfor-

mance. 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

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 predictions¹⁰. Laws have more impact on performance than conditions, since deterioration in results without laws was more significant than from models trained without condi-

tions. 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

¹⁰A 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.

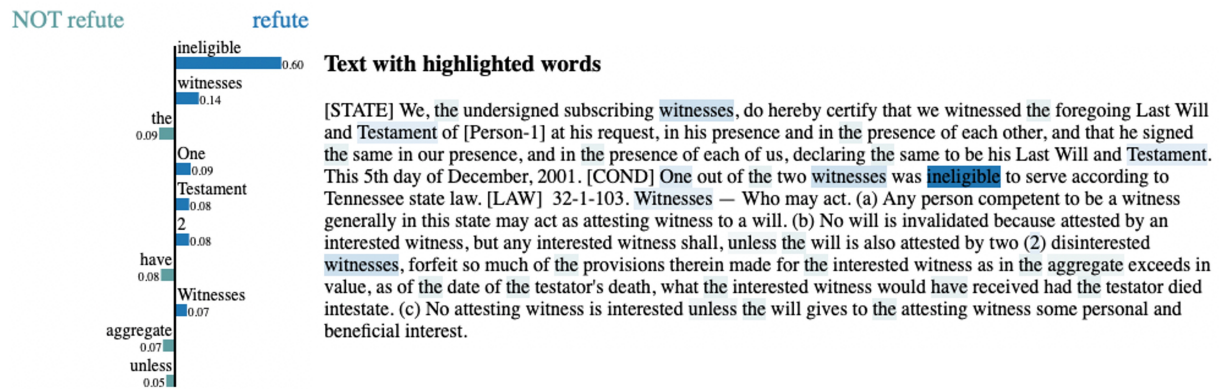


Figure 3: Examples of LIME explanations showing a case where the model (roberta-large-mnli) makes correct predictions based on features sensible to humans.

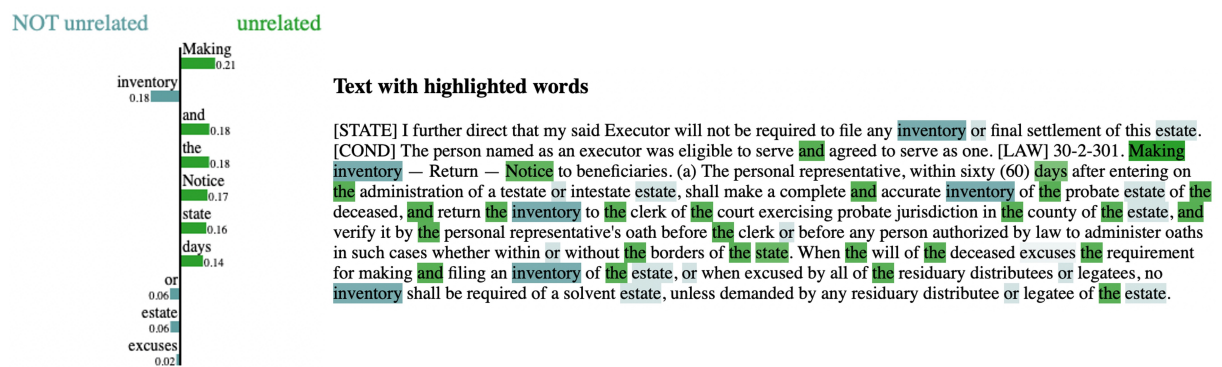


Figure 4: An example of a LIME explanation showing a case where the model (roberta-large-mnli) makes an incorrect prediction based on the tokens such as *Making*, *and*, and *the* which bear little semantic relevancy to the gist of the text.

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.

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 Sunawal, 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. [IFly-Legal: 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

1011C

arXiv > cs > arXiv:2210.16989v1

Search...

All fields

Search

Help | Advanced Search

Cornell University

We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors.

Donate

Computer Science > Computation and Language

[Submitted on 30 Oct 2022]

Validity Assessment of Legal Will Statements as Natural Language Inference

Alice Saebom Kwak, Jacob O. Israelsen, Clayton T. Morrison, Derek E. Bambauer, Mihai Surdeanu

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.

Comments: 10 pages, 4 figures; To be published in the Findings of the Association for Computational Linguistics: EMNLP 2022

Subjects: Computation and Language (cs.CL)

Cite as: arXiv:2210.16989 [cs.CL]
(or arXiv:2210.16989v1 [cs.CL] for this version)
<https://doi.org/10.48550/arXiv.2210.16989>

Submission history

From: Alice Kwak [view email]
[v1] Sun, 30 Oct 2022 23:53:13 UTC (2,195 KB)

Access Paper:

View PDF

TeX Source

CC BY-NC-SA

view license

Current browse context:

cs.CL

< prev

|

next >

new

|

recent

|

2022-10

Change to browse by:

cs

References & Citations

NASA ADS

Google Scholar

Semantic Scholar

Export BibTeX Citation

Bookmark

Exhibit 1026 Page 109

Submission history

From: Alice Kwak [\[view email\]](#)
[v1] Sun, 30 Oct 2022 23:53:13 UTC (2,195 KB)

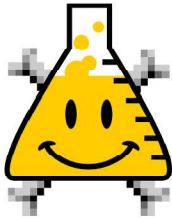
- Bibliographic Tools
- Code, Data, Media
- Demos
- Related Papers
- About arXivLabs

arXivLabs: experimental projects with community collaborators

arXivLabs is a framework that allows collaborators to develop and share new arXiv features directly on our website.

Both individuals and organizations that work with arXivLabs have embraced and accepted our values of openness, community, excellence, and user data privacy. arXiv is committed to these values and only works with partners that adhere to them.

Have an idea for a project that will add value for arXiv's community? [Learn more about arXivLabs.](#)



[Which authors of this paper are endorsers?](#) | [Disable MathJax](#) ([What is MathJax?](#))

← → ↺

web.archive.org/web/2022101102929/https://arxiv.org/abs/2210.16989v1

☆

Sign in

🔍

☰

INTERNET ARCHIVE

Wayback Machine

1 capture

1 Nov 2022

Go

OCT

NOV

DEC

2021

2022

2023

About this capture

 Cornell University

We gratefully acknowledge support from the Simons Foundation and member institutions.

arXiv > cs > arXiv:2210.16989v1

Search... All fields Search

Help | Advanced Search

Computer Science > Computation and Language

[Submitted on 30 Oct 2022]

Validity Assessment of Legal Will Statements as Natural Language Inference

Alice Saebom Kwak, Jacob O. Israelsen, Clayton T. Morrison, Derek E. Bambauer, Mihai Surdeanu

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.

Comments: 10 pages, 4 figures; To be published in the Findings of the Association for Computational Linguistics: EMNLP 2022

Subjects: **Computation and Language (cs.CL)**

Cite as: [arXiv:2210.16989](https://arxiv.org/abs/2210.16989) [cs.CL]
(or [arXiv:2210.16989v1](https://arxiv.org/abs/2210.16989v1) [cs.CL] for this version)
<https://doi.org/10.48550/arXiv.2210.16989>

Submission history

From: Alice Kwak [[view email](#)]
[v1] Sun, 30 Oct 2022 23:53:13 UTC (2,195 KB)

Download:

- PDF
- Other formats



Current browse context:
cs.CL
[< prev](#) | [next >](#)
[new](#) | [recent](#) | [2210](#)
Change to browse by:
cs

References & Citations

- NASA ADS
- Google Scholar
- Semantic Scholar

Export Bibtex Citation

Bookmark
    

Exhibit 1026 Page 111

Submission history

From: Alice Kwak [view email]
[v1] Sun, 30 Oct 2022 23:53:13 UTC (2,195 KB)

Bibliographic Tools

Code & Data

Demos

Related Papers

About arXivLabs

Bibliographic and Citation Tools

☐ Bibliographic Explorer (What is the Explorer?)

☐ Litmaps (What is Litmaps?)

☐ scite Smart Citations (What are Smart Citations?)

Which authors of this paper are endorsers? | Disable MathJax (What is MathJax?)

← → ↺ web.archive.org/web/20221101104607/https://arxiv.org/pdf/2210.16989v1.pdf ☆ Sign in

INTERNET ARCHIVE Wayback Machine <https://arxiv.org/pdf/2210.16989v1.pdf> Go 2 captures 1 Nov 2022 – 22 Apr 2023

OCT NOV DEC 01 2021 2022 2023 About this capture

1 of 10 Automatic Zoom

Validity Assessment of Legal Will Statements as Natural Language Inference

Alice Saebom Kwak, Jacob O. Israelson, Clayton T. Morrison,
Derek E. Bambauer, Mihai Surdeanu
The University of Arizona, Tucson, Arizona, USA
{alicekwak, jisraelson, 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

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

Oct 2022

Validity Assessment of Legal Will Statements as Natural Language Inference

Alice Saebom Kwak, Jacob O. Israelson, Clayton T. Morrison,
Derek E. Bambauer, Mihai Surdeanu

The University of Arizona, Tucson, Arizona, USA

{alicekwak, jisraelson, 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

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-

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

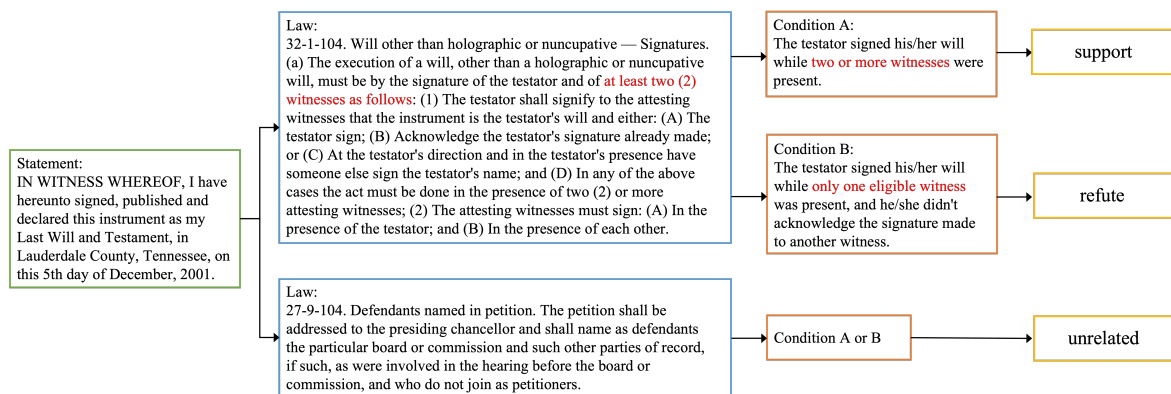


Figure 1: A demonstration of the annotation procedure. Given a law and condition, each statement is evaluated for validity. *Support*: it is supported by the law and condition provided; *Refute*: refuted by the law and condition. *Unrelated*: not relevant to the law. When a statement and a law are unrelated 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 OCR³, and copying statements into the dataset;

(2) for each statement, identifying the Tennessee

²A 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.

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

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

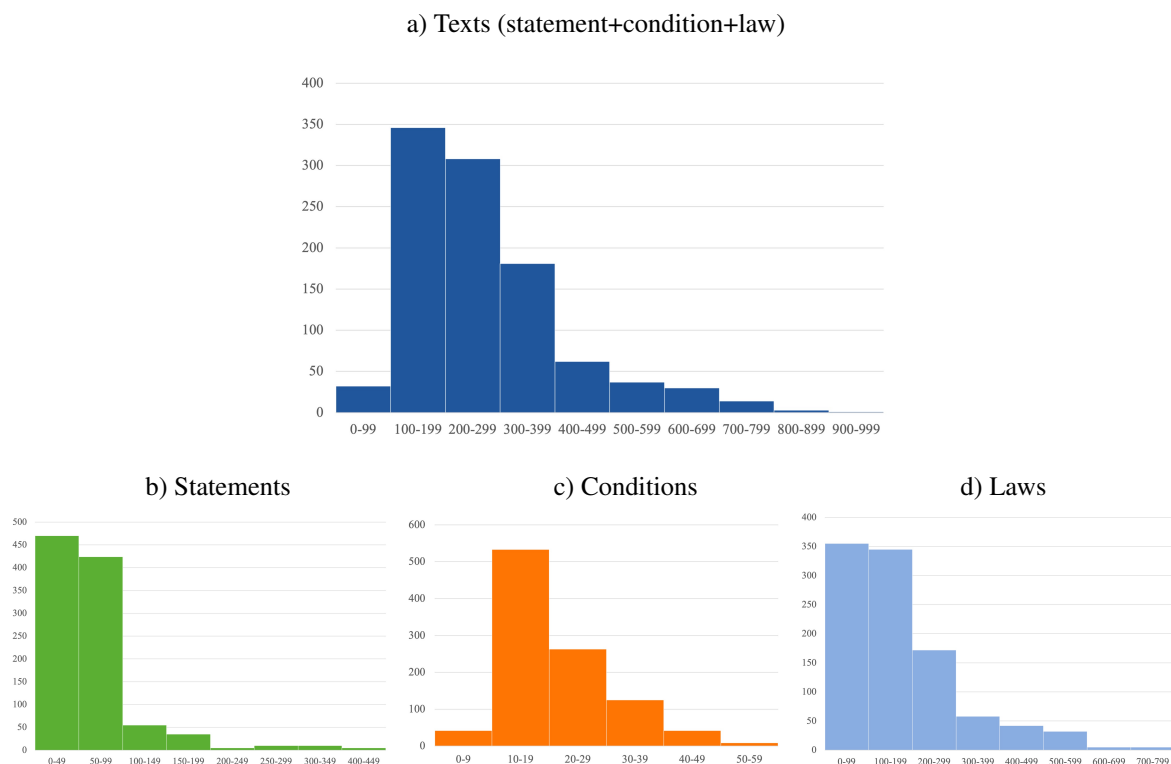


Figure 2: Histograms demonstrating the length distribution of a) full texts (i.e., statement+condition+law), b) statements, c) conditions, and d) laws in terms of token counts. Token counts are plotted on the X-axis; the Y-axis indicates the number of texts/statements/conditions/laws for each bin.

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 distilbert-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 page⁴) were chosen: all-mpnet-base-v2⁵, multi-qa-mpnet-base-

⁴<https://www.sbert.net>

⁵<https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

dot-v1⁶, 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

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>

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

⁹For 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.

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 perfor-

mance. 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

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 predictions¹⁰. Laws have more impact on performance than conditions, since deterioration in results without laws was more significant than from models trained without condi-

tions. 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

¹⁰A 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.

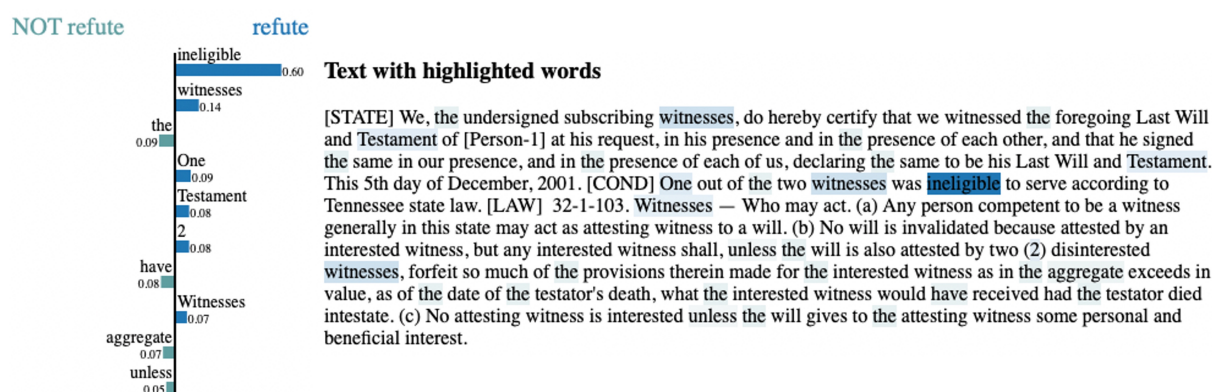


Figure 3: Examples of LIME explanations showing a case where the model (roberta-large-mnli) makes correct predictions based on features sensible to humans.

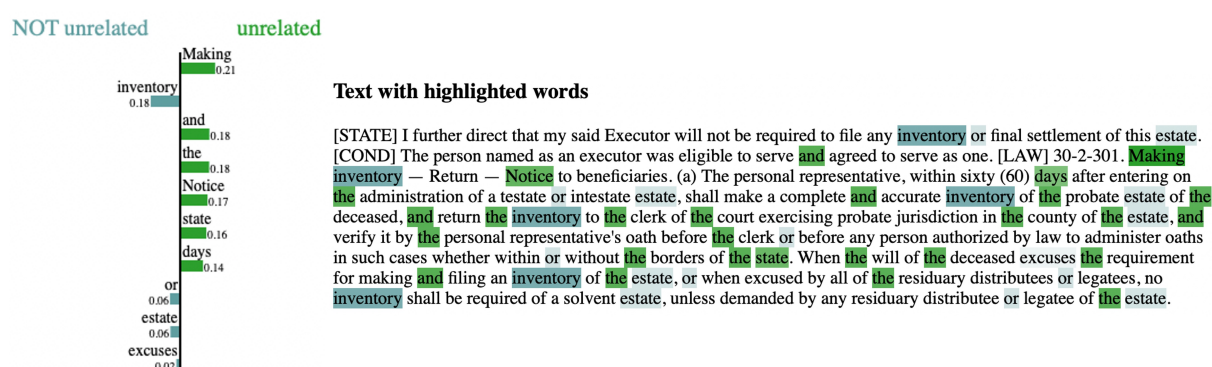


Figure 4: An example of a LIME explanation showing a case where the model (roberta-large-mnli) makes an incorrect prediction based on the tokens such as *Making*, *and*, and *the* which bear little semantic relevancy to the gist of the text.

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.

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 Sunawal, 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. [IFly-Legal: 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](#).