日本語判決書を用いたデータセットの構築

東京工業大学 情報理工学院
山田寛章
yamada.h.ax@m.titech.ac.jp

判決書とは？

- 判決書の特徴
  - 裁判の結果、**主文・理由**が記載してある文書
  - **長く複雑な文**から構成される(通常の文書の約2倍)
  - **複数の争点や階層からなる議論構造**
  - 文書長自体も長い
- 想定されるNLPのタスク
  - 固有表現抽出・匿名化(仮名化), 情報抽出, 議論マイニング, 自動要約, 機械翻訳, 文書検索, 法律QA, 判決予測, etc…

判決書を扱うデータセットの例 (海外)

- CAIL2018 (https://arxiv.org/abs/1807.02478)
  - 中国語, 約260万件, 刑事事件
  - データ収集元：中国最高人民法院
  - タスク
    - 入力：事実関係
    - 出力：適用法, 罪状, 刑期
- ECtHR dataset (https://www.aclweb.org/anthology/P19-1424/, ACL2019)
  - 英語, 約1万件, 人権侵害事件
  - データ収集元：European Court of Human Right (ECtHR)
  - タスク
    - 入力：事実関係
    - 出力：人権侵害の有無、違反条文の特定、事件の重要度判定
- New ECtHR dataset (https://arxiv.org/abs/2103.13084, NAACL2021)
  - 英語, 約1万件, 人権侵害事件, ECtHR datasetの拡張
  - タスク
    - 入力：事実関係
    - 出力：違反申立対象の条文+その理由となる事実関係の段落
- LexGLUE (https://arxiv.org/abs/2110.00976, NLLP2021@EMNLP2021)
  - 英語, 法分野特化版GLUE, 上記ECtHRや米国最高裁などのデータを含む。
  - タスク
    - ECtHR, SCOTUS, EUR-LEX, LEDGAR, UNFAIR-ToS, CaseHOLD

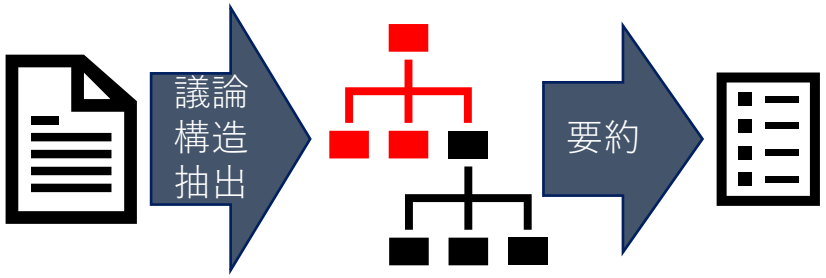
関連コンペ：COILEE

- 法律文書の情報抽出および含意関係認識を行う国際コンペティション
- カナダ連邦裁判所の判例データを用いた情報検索タスクと含意関係検索タスクの他、タスクの一部として日本司法試験短答式問題があり、データセットも提供されている

https://www.anlp.jp/proceedings/annual_meeting/2019/pdf_dir/E4-5.pdf など 毎年開催

構築済/中 データセット紹介

- JJD-WLコーパス*
  - 日本語民事事件下級審判決書コーパス (120 docs, 3.2M chars)
  - 人手による議論マイニングタスクのアノテーション付与済
  - 全文書について専門家作成の要旨を付与済

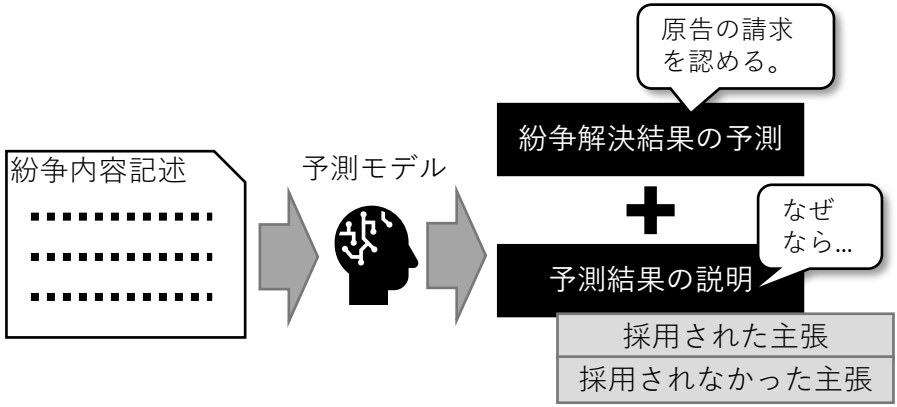


データ提供: ウエストロー・ジャパン株式会社

JSPS KAKENHI Grant Number JP 20J14385

*Hiroaki Yamada. 2021. Extracting argument structure from Japanese judgment documents for structure-based summarisation. Doctoral thesis.
*Hiroaki Yamada, Simone Teufel and Takenobu Tokunaga. 2019. Building a Corpus of Legal Argumentation in Japanese Judgement Documents: Towards Structure-Based Summarisation. Artificial Intelligence and Law. Springer Netherlands, 27(2):141–170.

- 日本語不法行為事件データセット（仮称）
  - 日本語民事事件第1審で、不法行為を取り扱う判決書で構成される(約4400件, 100M chars)
  - 47名のアノテータによって、不法行為の成否に関する裁判所の判断と、関連する主張・証拠がアノテーション済



データ提供: 株式会社 LIC

JST ACT-X Grant Number JPMJAX20AM (代表：山田寛章/東工大)
JST RISTEX Grant Number JPMJRX19H3 (代表：角田美穂子/一橋大)

本日のトピック

- データの入手性
  - ライセンス・権利関係
  - データ取得先の確保
  - データ形式リスク
- アノテーション
  - アノテーションスキームの開発
  - 一致度測定
  - 作業者の確保
  - 作業者の訓練＋大規模作業の展開

判決書の入手性と壁

- **判決文自体には著作権はない**（著作権法１３条３号）
- 裁判所Webサイトに掲載されているPDFファイルをDLして使う分には問題なし。ただし…
  - 網羅的ではない
    - Cf: 米国Caselaw Access Project, 欧州人権裁判所European Court of Human Rights
  - 裁判所が「重要」とした事件を中心として収録されているため、バイアスが存在
    - 判決予測システムの訓練などでは問題になる
  - HTMLやXMLではなくPDF→処理しやすい形に変換する手間**大**
    - 定型的な表現をヒントにパーザーを開発して、節や段落を構造化・抽出する研究もある: Igari, Hirokazu, Akira Shimazu, and Koichiro Ochimizu. 2012. “Document Structure Analysis with Syntactic Model and Parsers: Application to Legal Judgments.” In , 126–40. Springer, Berlin, Heidelberg.

  →実際には判例データベースを販売している出版社などを頼ることになる
- **判例データベースは「データベースの著作物」**として著作権による保護の対象となり得る
  - 使用には出版社との交渉や契約の締結が前提
- **オープンデータ化で入手性改善に希望が！**

判決書の潜在的なリスク

- 機密情報・個人情報・プライバシー情報保護の懸念
  - 前提：
    - 裁判の公開は憲法の原則（日本国憲法第八十二条）
    - 訴訟記録の閲覧も可能（民事訴訟法第九十一条）
    - 当事者は「私生活についての重大な秘密」「営業秘密」が記録されていることを理由として閲覧の制限を申し立てが可能（民事訴訟法第九十二条）
    - 公開済PDFや判例DB上では、固有表現（人名・地名・会社名など）は多くが仮名化されている

    →最低限の保護は存在する
  - それでも・・・
    - データセットの構築後・公開後に非公開にするよう要求があったら？
    - 仮名化の抜け・漏れによるプライバシー情報漏洩？
    - 完全な匿名化ではなく、固有表現の置換による仮名化処理では、当事者の特定を完全に防ぐことは困難

→データセット構築後・配布後の追跡やアップデートのため、配布対象の連絡先の把握や、再配布禁止など、ライセンスや配布の仕組み上の工夫が必要な可能性がある

アノテーション作業の展開

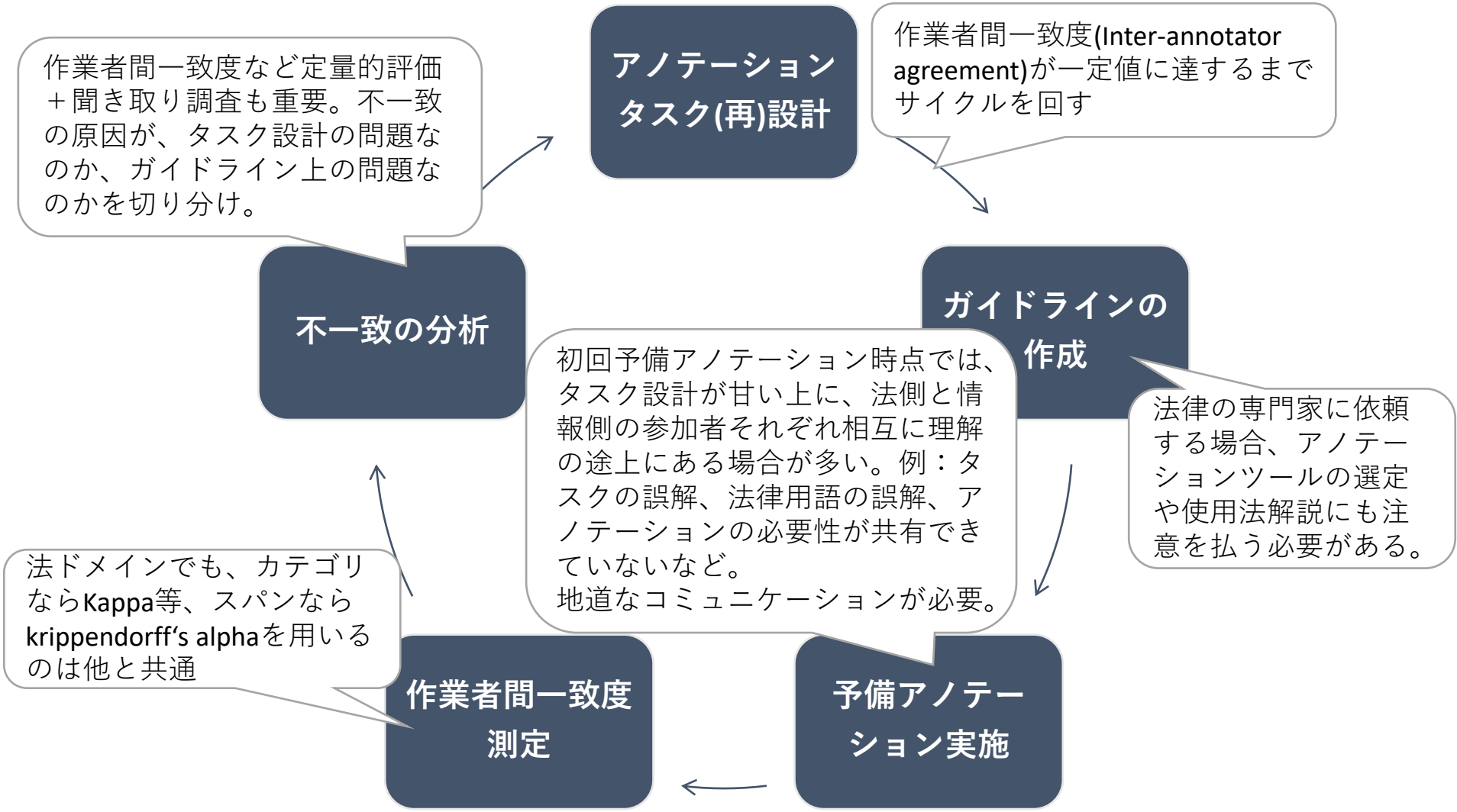
1. データセットが対象とするタスクの設計
2. データの選定
3. 訓練・評価に必要なアノテーションの決定
4. **アノテーションスキームの開発**
5. **本アノテーション実施**
6. アノテーションデータの回収

「専門的知識」を持つアノテータの確保

- クラウドソーシングは×
  - 判決書のように、読むこと自体に専門的知識を要する文書のアノテーションとは相性が悪く、必要な人材を確保しにくい。
- アノテーション会社に委託？
  - アノテーションの専門家はいても、法的知識も兼ね備えることは希
- ベスト：法学研究者や法曹
  - 知り合うことが難しい！＋NLPに興味・理解のある人は少ない

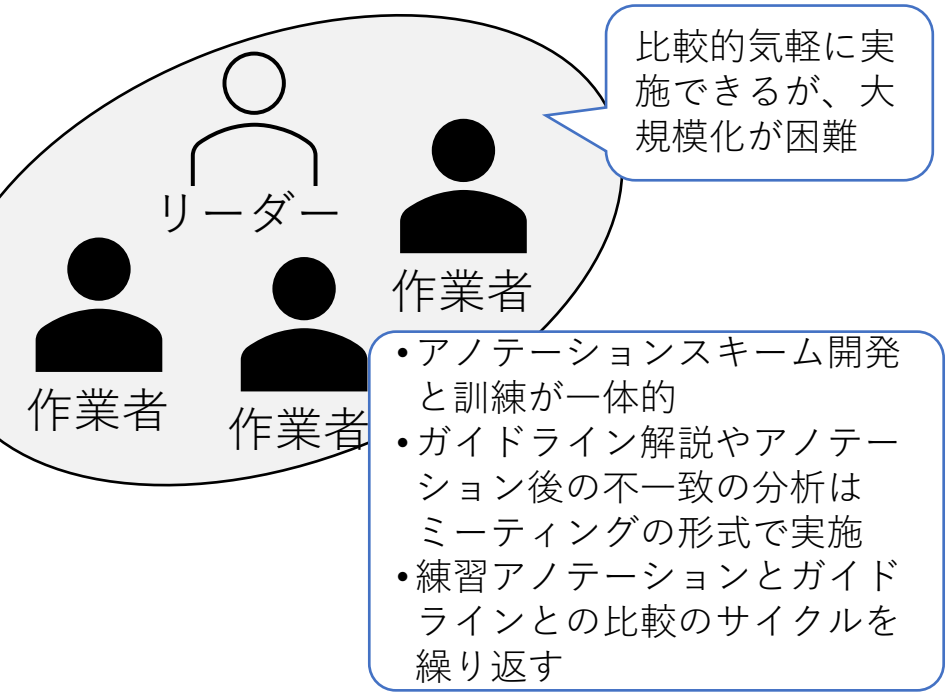
  →伝手を頼るしかない･･･人脈は宝

アノテーションスキーム開発

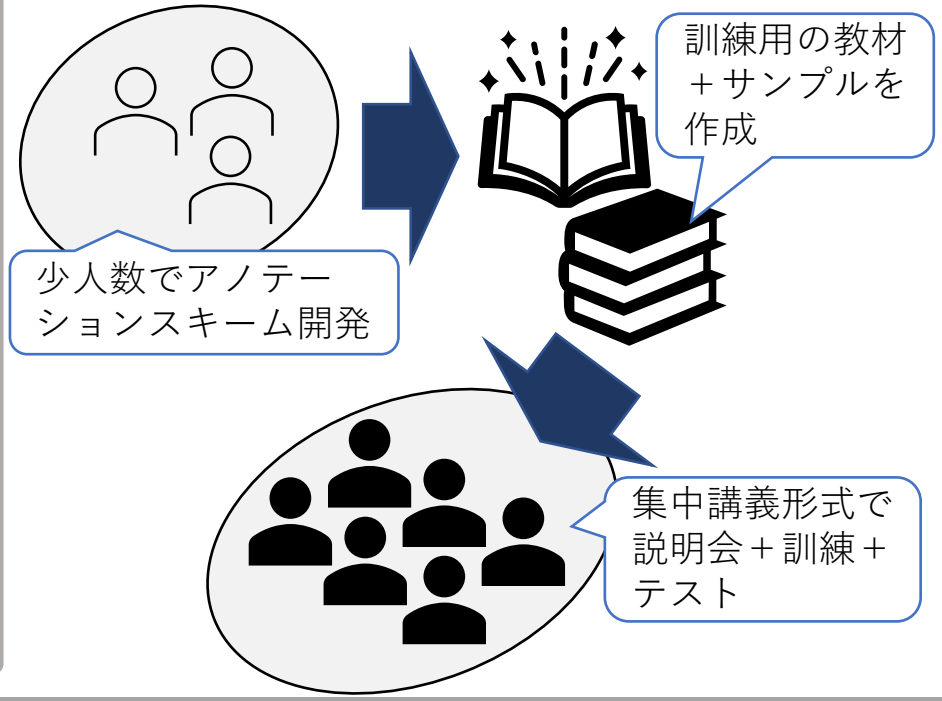


アノテータ訓練の例

- **JJD-WLコーパス**
  - アノテータ：2人 (大学院生/法学研究科など)
  - アノテーションタスク：
    - 議論的テキスト抽出(スパン)＋スパン間の支持関係特定(リンク)など
  - 訓練方法：ミーティング型

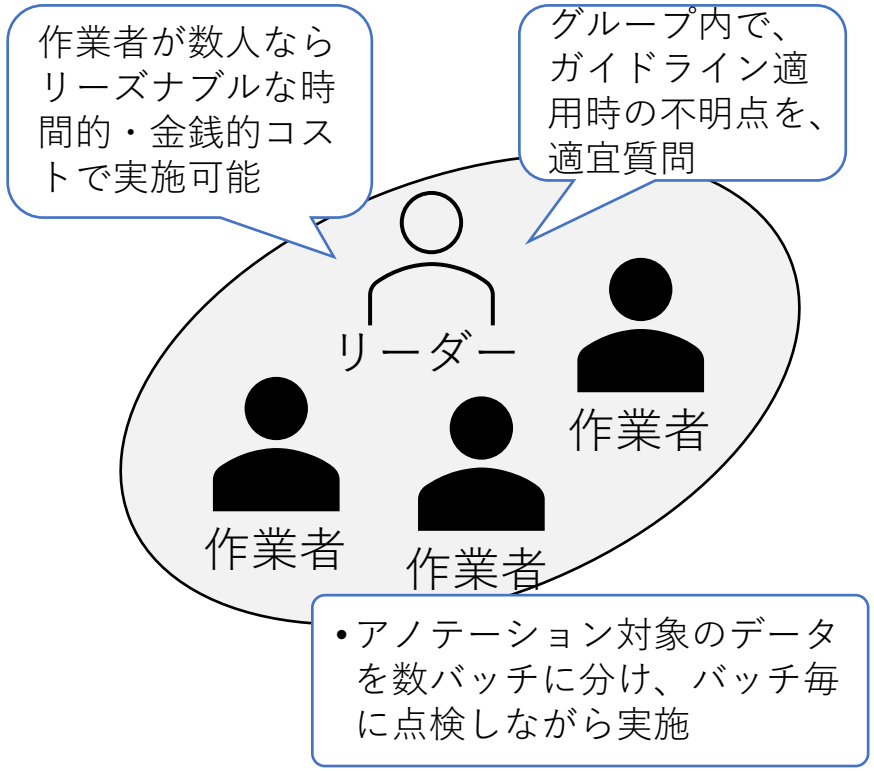


- **不法行為事件データセット**
  - アノテータ：47人 (法曹・司法試験合格者・法科大学院学生など)
  - アノテーションタスク：
    - 不法行為判断の抽出＋関連する主張の抽出（スパン）＋各主張の採否特定など
  - 訓練方法：セミナー型

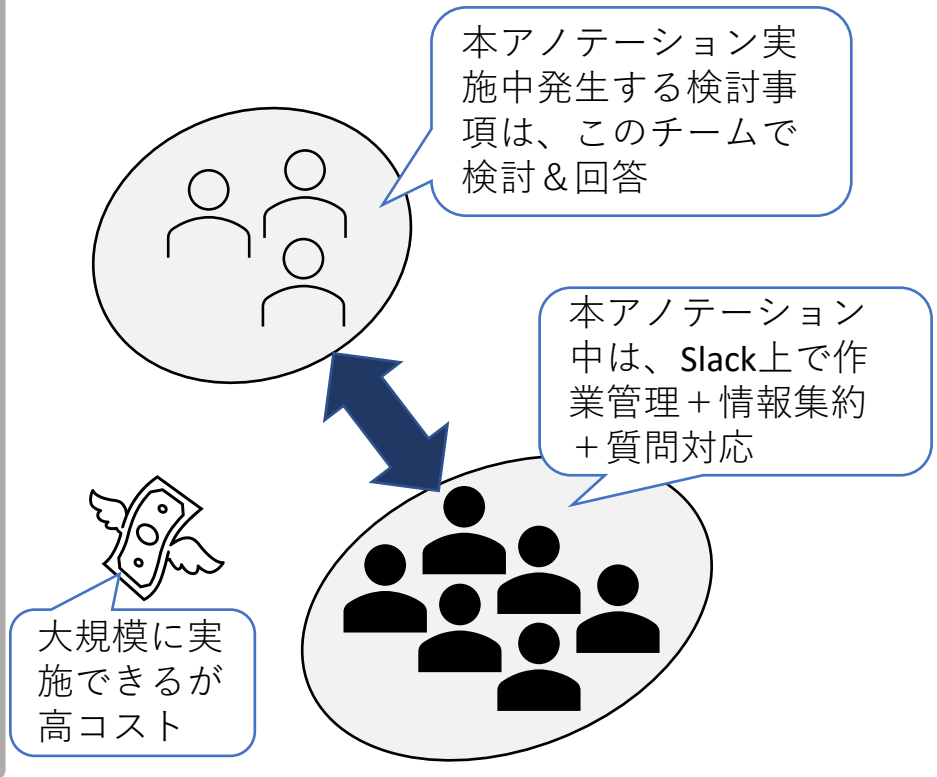


本アノテーション実施の例

- **JJD-WLコーパス**
  - アノテータ：2人 (大学院生/法学研究科など)
  - アノテーションタスク：
    - 議論的テキスト抽出(スパン)＋スパン間の支持関係特定(リンク)など



- **不法行為事件データセット**
  - アノテータ：47人 (法曹・司法試験合格者・法科大学院学生など)
  - アノテーションタスク：
    - 不法行為判断の抽出＋関連する主張の抽出（スパン）＋各主張の採否特定など



まとめ

- 日本語・日本法の判決書を用いて、深層学習手法の訓練・テストにも耐える規模のデータセットを作成中。
- 判決書特有の潜在的リスクを念頭に、データセットの構築と共有枠組みを検討する必要がある。
- 判決書のアノテーションには、法学系研究者や法曹などの法律実務家との連携が重要。
- 専門的なアノテーションの大規模実施は、時間と人材と資金を揃える必要があるのでハードルが高いが、理想的には実施すべき。

今後の展開

- 日本語不法行為事件データセット（仮称）は次年度中に公開または共有を開始する予定
  - ライセンスや公開の形態は目下検討中
- 民事事件判決のオープンデータ化や関連する司法データの公開・オープン化の状況を注視
  - データが入手できることが第一歩
- **法分野テキストを用いたNLP研究をやってみたい人募集中**