学会記事

Non Peer-Reviewed

日本語における評価用データセットの構築と利用性の向上 —JED2022 ワークショップの成果と展望

松田　寛[†]・柴田　知秀[††]・河原　大輔[†††]・久本　空海[††††]・
久保　隆宏[†††††]・浅原　正幸[††††††]

1　ワークショップ開催の経緯と構成

自然言語処理研究を目的とする日本語の大規模テキストデータセットの構築・公開は 1990 年代に入って本格化し，既に 1995 年の言語処理学会第 1 回年次大会では，辞書・コーパスに関する 4 つのセッションで各種データセットの構築と利活用について多くの発表があった．この時期の大規模テキストデータセットは元テキストとして新聞記事を使うものが多く，そうしたデータセットの利用には CDROM の購入や個別のライセンス契約が必要な場合があり，大規模テキストデータセットで訓練を行った機械学習モデルの配布・利用の障壁となる場合があった．

2000 年代になって，急速なインターネットの普及により，Email や SNS・Web コンテンツに由来する大量のテキストが容易に入手可能となった．こうしたインターネット由来のテキストを自ら収集して研究用途で利用するのはライセンス面の制約が少ないため，様々な応用タスクに向けたデータセットの構築が進んでいった．しかし，インターネット由来のテキストについて，当時の日本の著作権法では，閲覧以外の目的で利用可能な範囲が明確でなく，また研究目的で比較的自由に著作物を利用することが可能ないわゆる「フェアユース」の規定もなかったため，インターネット由来のテキストを含むデータセットの公開時には，利用許諾に「商用利用の禁止」や「要請に基づく削除義務」のような条項を含める必要が生じていた．

2010 年代に入ると，自然言語処理技術の性能を評価するための複数のタスクを束ねた英語向けの評価用データセットが，無償かつ商用利用も認める標準的な OSS（オープンソース・ソフトウェア）ライセンスやオープンデータライセンスのもとで数多く公開され，広く利用されていった．利用者は GitHub リポジトリからデータセットと評価ツール類をクローンするだけで，手軽に実験を行うことが可能になった．こうした英語のデータセットのテキストを日本語に翻訳する試みもあったが，翻訳プロセスに由来するアーティファクトや文化の相違に起因するバイア

† 株式会社リクルート Megagon Labs
†† ヤフー株式会社
††† 早稲田大学
†††† 株式会社 MIERUNE
††††† アマゾンウェブサービスジャパン合同会社
†††††† 国立国語研究所・東京外国語大学

スによる品質劣化が顕著に見られたため，日本語のテキストをオリジンとする評価タスクの構築が必要であることが認識されていった．やがて，最初から日本語テキストで構築した各種評価タスクのデータセットの開発・公開が活発になっていったが，それらの成果を束ねて GLUE のような評価用データセットとして公開する動きはなかなか広まらなかった．その要因として，トップ会議を目指す論文になりにくいこと，評価タスクを追加するためデータセットを一から構築するのに大きなコストがかかることに加えて，次の法務リスクが意識されていたと考えられる．ある評価用データセットに含まれる評価タスクのライセンスに，独自に定義されたものが一部でも含まれていると，評価用データセットの利用者は弁護士や法務部門の協力を受けながら，ライセンスの条文を個別に精査する必要が生じるため，こうしたコストをかけられる利用者は限定されてしまい，利用の広がりが見通せない中では評価用データセットを公開する機運は高まりにくかった．やがて，ELMo や BERT をはじめとする事前学習モデルの性能が各種 NLP タスクを席巻する時代になってからは，日本語評価用データセットへのニーズは益々顕在化していった．

このような課題認識のもとで，NLP2022 で開催した「日本語における評価用データセットの構築と利用性の向上」ワークショップ (JED2022) では，日本語における評価用データセットの構築手法そのものに加えて，データセットの公開方式・利用性の高いライセンス・タスクの複合化といった応用を容易にするための取り組みを集めて議論することで，日本語データセット公開の流れを加速し，日本語 NLP 業界全体のさらなる発展および生産性向上につながる議論の場となることを目指した．また，著作権法など評価用データセット公開時に考慮すべき法令に関する特別講演を弁護士の柿沼太一先生に依頼した．ワークショップの構成は次の通り（詳細は JED2022 プログラムページ[1]を参照）．

特別講演　日本語データセットの構築・利用・公開に関する法的整理
口頭発表　日本語評価用データセットの構築と公開（14 件）
テーマセッション　リーガル分野におけるデータセット構築・利活用の現状と展望（4 件）
リレートーク　形態素解析辞書・統語解析データセットの公開
総合討論　日本語評価用データセットとその学習モデルが協調的な発展を遂げるために

2　特別講演

本ワークショップでは，STORIA 法律事務所の柿沼太一先生をお招きし，特別講演「日本語データセットの構築・利用・公開に関する法的整理」を企画した．当該時間には 300 名ほどの聴講者が集まり，大変盛況であった．

[1] https://jedworkshop.github.io/jed2022/program/

データセットの構築・利用・共有においては，法令上の制限や契約・ライセンスなどを検討する必要がある．法的な整理として，2018 年以降 4 回の改正が行われた著作権法や 2022 年 4 月改正の個人情報保護法の動向がデータ構築者における関心事項である．さらには，データセット構築者やデータセットを利用して深層学習等のモデルを構築する際のライセンスの扱いについても様々な検討が必要である．

今回，柿沼先生に問題の全体像についてご説明いただき，さらに多様な具体事例における検討事項について解説していただいた．特に，「個々のデータに関する知的財産権（著作権等）」と「データの集積行為に関する知的財産権（DB 著作権及び限定提供データ）」と「契約・ライセンス」の組み合わせについて整理を行っていただいた．

3 日本語評価用データセットの構築と公開

「日本語評価用データセットの構築と公開」では午前に 2 セッション，午後に 1 セッションの計 3 セッションを行い，3 セッションあわせて 7 件の一般発表と 7 件のライトニングトーク，計 14 件の発表があった．評価用データセットに関するワークショップは初めてであったので，ワークショップを開催できるほどの発表が集まるか不安であったが，多数の発表申し込みをいただけた．

日本語評価用データセットを構築されている方からはデータセットの内容や構築手順だけでなく，クラウドソーシングなどを利用したデータ収集方法や言語資源公開のノウハウなどもご発表いただき，論文には書かれにくい知見・苦労・裏話などが共有された．その他にも既存のデータセットに関する詳細な分析や，事前学習モデル構築に関する知見の発表もあった．

以下に発表を列挙する．タイトルの末尾に ↯ が付与されている発表はライトニングトーク，その他は一般発表を示す．

日本語評価用データセットの構築と公開 (1)

- JGLUE: 日本語言語理解ベンチマーク
  柴田知秀（ヤフー）, 栗原健太郎, 河原大輔（早大）
- 日本語版 CoLA の構築の舞台裏
  染谷大河, 大関洋平（東大）
- QA における評価用データセットの役割と日本語 QA データセットの必要性についての考察
  田保健士郎, 小林景（慶應大）

日本語評価用データセットの構築と公開 (2)

- 闘病ブログ記事に対する投与薬剤の奏功情報アノテーション ↯
  高山隼矢, 荒瀬由紀（阪大）, 梶原智之（愛媛大）, Chenhui Chu （京大）

- WRIME：主観と客観の感情極性分類のための日本語データセット ↯
  鈴木陽也, 宮内裕人, 秋山和輝, 梶原智之, 二宮崇（愛媛大）, 武村紀子, 中島悠太, 長原一（阪大）
- 学習データセット改善によるアスペクト感情分析モデルの性能改善
  亀谷聡（インテック）
- 日本語レシピデータセットの継続的な構築と複合的な利用
  原島純, 平松淳, 深澤祐援, 山口泰弘（クックパッド）

日本語評価用データセットの構築と公開 (3)

- 供述調書に現れる数量表現の推論テストセットの構築
  小谷野華那（お茶大）, 谷中瞳（東大）, 峯島宏次（慶應大）, 福田浩司, 橋爪宏典 (NEC), 戸次大介（お茶大）
- JaNLI: 日本語の言語現象に基づく敵対的推論データセット
  谷中瞳（東大）, 峯島宏次（慶應大）
- Japanese Realistic Textual Entailment Corpus の紹介 ↯
  林部祐太 (Megagon Labs)
- SNS を出典とする言語資源の公開にまつわるノウハウ ↯
  榊剛史, 水木栄（ホットリンク）
- クラウドソーシングに基づく日本語タスク指向型対話収集基盤の構築に向けて ↯
  邊土名朝飛, 友松祐太 (AI Shift), 阿部香央莉, 佐々木翔大, 乾健太郎（東北大学）
- 日本語転移学習モデルにおける事前学習コーパスのフィルタリング ↯
  渡邊亞椰, 河原大輔（早大）
- 公開日本語言語モデルとその評価の現状 ↯
  林政義（ワークス）

今回一度だけのワークショップではなく，今後も継続的にこのような機会を設け，日本語評価用データセットに関する議論が進められればと考えている．

4 リーガル分野におけるデータセット構築・利活用の現状と展望

リーガル分野（法領域）では，テキスト情報が比較的厳密な形の文章で記述されるため，自然言語処理技術との相性は良いと想定され，その活用による社会への影響も大きいと考えられるが，日本においては，そうした取り組みは研究・産業いずれにおいても未だ活発ではない．その大きな要因として，データセットの公開・整備が十分なされていないことが挙げられる．

本セッションでは，当分野におけるデータセットの構築とその利活用に取り組んでいる企業・大学の研究者から，取り組みの背景と現状，知見や今後の課題を共有いただいた．また，発表

後にはパネルディスカッションを設け，聴講者からも質問を募り議論した．

まず，株式会社 Legalscape の八木田樹氏から，企業内の法務関連の情報の大部分が非公開，かつ，機械可読性の低い状態である現状の解説の後に，この数年で動き始めた政府によるオープンデータ化の取り組みが紹介された．それらを踏まえ，法情報の民主化に伴う大変革へ向けて重要なのは「法情報をすべて整理し，インフラ的に提供すること」との主張が述べられた．

次に，MNTSQ 株式会社の稲村和樹氏から，フェアな合意の実現へ向けた契約書解析と，そのためのデータセット構築への取り組みを解説いただいた．そこでの課題として，データの分散や閉塞といった組織ガバナンスも含めたハードルの存在，類型や業種ドメインのカバレッジ，専門家知見が不可欠なことによるアノテーションコストの高さ，などが紹介された．

続いて，株式会社 LegalForce の舟木類佳氏から，法務家による契約書の理解支援のための，権利義務の自動認識を目的としたコーパスの構築について解説いただいた．このコーパスには，当事者・権利・義務・要件・例外に関する範囲情報と，各情報同士の関連情報がアノテーションされている．それらアノテーションのガイドラインや，アノテーション実施の際の困難な事例が共有された．

学術界からは，東京工業大学の山田寛章先生から，日本語判決書を用いたデータセットについて報告いただいた．海外に比べ，日本語で書かれた日本法に対応するデータセットの普及は進んでいない．現状を踏まえ，過去に構築した判決書議論マイニングデータセットおよび，現在構築中の判決書判断予測データセットについて，概要と構築過程が紹介された．特に，法律ドメインにおけるアノテーションについて，実際の構築作業を通じて得られた知見が共有された．

最後のパネルディスカッションでは，参加者からも多数の質問が寄せられた．アノテーション業務の実情や，分野特有のアプローチ，国外での事例などの議論が活発に起こり，セッション終了後もオンライン上で続いた．自然言語処理にとってまだ馴染みの薄い「法」という領域について，実際に取り組む者が集い共有する貴重な場になったと考える．

5 形態素解析辞書・統語解析データセットの公開

本セッションでは，形態素解析辞書及び係り受けアノテーションデータ整備について議論した．日本語の分かち書きの単位と深層学習モデルの関係，統語情報アノテーションが主に話題となった．

深層学習においては訓練時に単語分かち書きのみを用いる傾向にあるが，産業応用においては単語に紐づく様々な情報が活用されている．ワークスアプリケーションズ・システムズ社の Sudachi の同義語辞書や，その単位に基づく単語埋め込み・事前学習モデルなどが紹介された．

係り受けアノテーションデータ整備については，日本語の伝統的な文節係り受け木から，多言語で共有可能な係り受けアノテーションフレームワーク Universal Dependencies への変換につ

いて議論した．形態に基づく単位である国語研短単位と品詞の用法まで同定した単位である国語研長単位の 2 種類の日本語 Universal Dependencies の 2 種類のデータの差異について紹介された．また日本語 Universal Dependencies に基づく依存構造解析モデルの構築について，spaCy および GiNZA での取り組みについて紹介された．

6 ワークショップの運営

ワークショップの開催にあたっては JED2022 公開ページ[2]を作成し，募集要項・新着情報の告知や，プログラムのタイムテーブル，各発表の題名・概要を公開する場として利用した．加えて，外部公開を許諾していただいた発表者のスライド資料もホスティングしている．ホームページはワークショップ閉会後も継続して公開状態を維持している[3]．このように誰もがアクセスできる包括的な公式の情報源を恒久的に提供することは，ワークショップの参加者・発表者，そして当日は参加できなかった未来の研究者らにとっても利便性があるものだと考える．

ワークショップ運営メンバーのコミュニケーションには，年次大会でも使用された Slack を用いた． 発表者は運営メンバーと同じ Slack に作成した専用のチャンネルへ招待し，当日までの依頼事項や発表方法などの案内を行った．発表者から発表資料を提出頂くにも Slack を用いた．なお，運営メンバーのチャンネルは発表者が参照できないよう運営メンバー限定とした．ワークショップに関わるコミュニケーションを Slack へ一本化することで，円滑な運営が行えた．年次大会の Slack は大会終了後消えてしまうが，ワークショップの Slack は継続する．発表者に連絡をしたタイミングや問い合わせ内容を参照できることは，来年の運営を改善する際にあたり有用である．

発表者に提出頂くのは論文ではなく発表スライドとした．本ワークショップは研究者同士で論文では書けない苦労話やノウハウを共有することで，言語資源構築の生産性を向上させることを企図しているためである．SNS を出典とする言語資源の公開にまつわるノウハウの発表ではデータを公開するために社内を説得する方法，また QA における評価用データセットの発表では単純な翻訳で日本語データセットを作る際の問題点が指摘されるなど，ワークショップが企図した苦労話やノウハウの共有が行えたと考えている．

ワークショップの前に発表者にワークショップで使用する Zoom の接続確認を頂いた．接続確認の時間で簡単ながら運営メンバーと発表者が交流できる機会を作ることができた．ともすると直接話すことがなく終わってしまうところ，リモートではありつつもカメラをオンにして対面で話せる機会があったことで，ワークショップの趣旨に賛同いただいた研究者の方と親睦

[2] https://jedworkshop.github.io/jed2022/
[3] ホームページのホスティングには GitHub Pages を利用している．

を深めることができた.

データセットとモデルが協調的に発展する未来にむけて，発表者の方に事前アンケートを行いアンケート結果についてワークショップの最後に議論を行った．どのようなタスクのデータセットを構築するかについては実用化のニーズを意識するという回答が見られた一方，データセットの活用促進や公開後の発展方法については特になしという回答が目立った．現状，公開後の活用，また更新について計画が立てられているケースが少ないことは協調的な発展を目指すにあたり課題といえる.

7 今後の展望

2022 年 9 月 7 日（水） に，JED2022 の一部の講演者のほか，言語資源を構築している若手研究者を招待し，「日本語における評価用データセットの構築と利用性の向上」分科会[4]を，国立国語研究所においてハイブリッド形式で開催する予定である．また，今後もワークショップを定期的に開催する予定であり，特に NLP2023 に向けては，日本語データセットと日本語事前学習モデルの相補的な発展を議論する場を設けたいと考えている.

謝 辞

JED2022 の開催にあたり，言語処理学会第 28 回年次大会大会委員会の皆様，ワークショップスポンサー 4 団体の皆様には，多大なご支援を賜りました．柿沼弁護士には，ご多忙の中，二度に渡りご講演の趣旨を詳細に議論する場を設けていただきました．紙面の都合で本稿の執筆を依頼できなかったワークショップ提案者の高岡一馬氏・林部祐太氏のご貢献に敬意を評します．最後に，ワークショップの発表者・参加者・スタッフの皆様に，厚くお礼申し上げます.

略歴

松田　寛：株式会社リクルート Megagon Labs 研究員・言語処理学会理事.
柴田　知秀：ヤフー株式会社上席研究員・言語処理学会理事.
河原　大輔：早稲田大学教授.
久本　空海：株式会社 MIERUNE ソフトウェアエンジニア.
久保　隆宏：アマゾンウェブサービスジャパン合同会社.
浅原　正幸：国立国語研究所・東京外国語大学教授.

[4] https://masayu-a.github.io/ELW/ELW2022/JED.html