学会記事

Non Peer-Reviewed

日本語言語資源の構築と利用性の向上 —JLR2023 ワークショップ

浅原　正幸[†]・河原　大輔[††]・久保　隆宏[†††]・坂口　慶祐[††††]・
柴田　知秀[†††††]・松田　　寛[††††††]

1　はじめに

我々は言語処理学会第 29 回年次大会にて併設ワークショップ「日本語言語資源の構築と利用性の向上」を企画開催した．本ワークショップは第 28 回年次大会の「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」に引き続き，2 回目の開催である．今回から後述するように事前学習モデルを含め，広く言語資源の構築と利用性を募集対象とした．前回は残念ながらオンライン開催であったが，今回はハイブリッド開催で，多くの発表者・聴講者と対面で交流できた．様々な分野の方々に参加いただき，会場内外で活発なディスカッションが行われた．

近年の言語処理においては，辞書やテキスト・音声・マルチモーダルコーパスのみならず，事前学習モデルが研究に必要な状況である．本ワークショップでは，事前学習モデルも言語処理における重要な言語資源と位置づけ，各種モデルの構築や利用についても発表を募集した．結果，一般発表 10 件とライトニングトーク 7 件の発表が行われた．

2　当日の様子

以下，各セッションごとの発表を簡単に紹介する．

「日本語データセットの構築」のセッションでは 4 件の一般発表と 2 件のライトニングトークが発表された．新納氏は，カナ漢字混じりのテキストに対して読みを付与するタスクのためのデータセット構築について発表した．異なる読みで実体が異なるもの・同じものなどの整理のほか，連濁の扱いの難しさについて紹介された．柴田は，日本語言語理解ベンチマーク JGLUE の整備状況について報告した．文書分類・文書ペア分類・質問応答の構築報告のほか，Leaderboards への掲載についても議論された．浅原は，日本経済新聞社の協力で 2023 年 3 月 13 日に公開さ

† 国立国語研究所（注：本記事執筆に際し，全著者は同等に貢献しました．）
†† 早稲田大学
††† アマゾンウェブサービスジャパン合同会社
†††† 東北大学
††††† ヤフー株式会社
†††††† 株式会社リクルート Megagon Labs

れた「日本経済新聞記事オープンコーパス」について発表した．現在までに進めた形態・統語情報アノテーションのほか，今後進める意味情報アノテーション・被験者実験による言語のとらえ方の収集について紹介した．富田氏は，CCG に基づくツリーバンク構築の試みとして，項構造が付与された ABC ツリーバンクに対して詳細な統語素性を付与する，リフォージングの取組について発表した．続いて戸次氏は，日本語 CCGBank の言語学的な妥当性について紹介した．特に日本語の受動・使役の現象について議論した．最後に Yin Yue 氏は，大規模な日本語音声コーパス ReazonSpeech の構築と，同データに基づく学習済み音声認識モデルについて発表した．

「多言語・多分野の言語資源の構築」セッションでは，3 件の一般発表と 3 件のライトニングトークが発表された．砂岡氏は，Code Switching による多言語混在言語資源の重要性について発表した．多言語混在言語データの音声認識や機械翻訳について各種ツールの比較を行った．宮本氏は，中日対訳コーパスの構築および著作権処理について発表した．ディスカッションではコーパスの著作権処理について活発な議論がされた．武田氏は，教育研究・実践のための言語資源・自然言語処理にあり方について発表した．内藤氏は，コペンハーゲンからオンラインで，デンマークの言語資源の構築手法について発表した．土肥氏は，難病・希少疾患の症例報告に基づく言語資源構築について報告した．橋田氏は，高等学校の授業の一環としてグラフ文書を構築する取り組みについて報告した．このように多言語または様々な分野における言語資源の取り組みが紹介され，言語資源構築の広がりが確認された．

招待講演として 株式会社 Studio Ousia の山田育矢先生に「知識拡張型言語モデル LUKE」[1] というタイトルでご講演いただいた．LUKE (Yamada et al. 2020)，多言語版 LUKE (Ri et al. 2022)，LUKE を用いたエンティティリンキング (Yamada et al. 2022; Oba et al. 2022)，日本語版 LUKE についてご講演いただいた．

「事前学習モデルの構築と利用」セッションにおいては，3 件の一般発表と 2 件のライトニングトークが発表された．塚越氏は，実験プログラムの言語資源化について発表した．この取り組みは同じ『自然言語処理』Vol. 30 No. 2 の別の学会記事にて詳しく紹介される．近藤氏は，日本語 BigBird の構築について，事前学習時におきたトラブルとその対処方法について発表した．小林氏は，日本語 DistilBERT[2] の構築とともに，LINE NLP チームの様々な公開ツールについて紹介した．植田氏は，日本語 DeBERTa モデル[3]の構築について紹介した．佐藤氏は，LINE NLP チームが整備している日本語事前学習モデル HyperCLOVA 構築時と，実用に向けて様々なサブシステム実装の必要性について発表した．

[1] https://speakerdeck.com/ikuyamada/zhi-shi-kuo-zhang-xing-yan-yu-moderuluke にて発表資料を公開.
[2] https://engineering.linecorp.com/ja/blog/line-distilbert-high-performance-fast-lightweight-japanese-language-model
[3] https://huggingface.co/ku-nlp

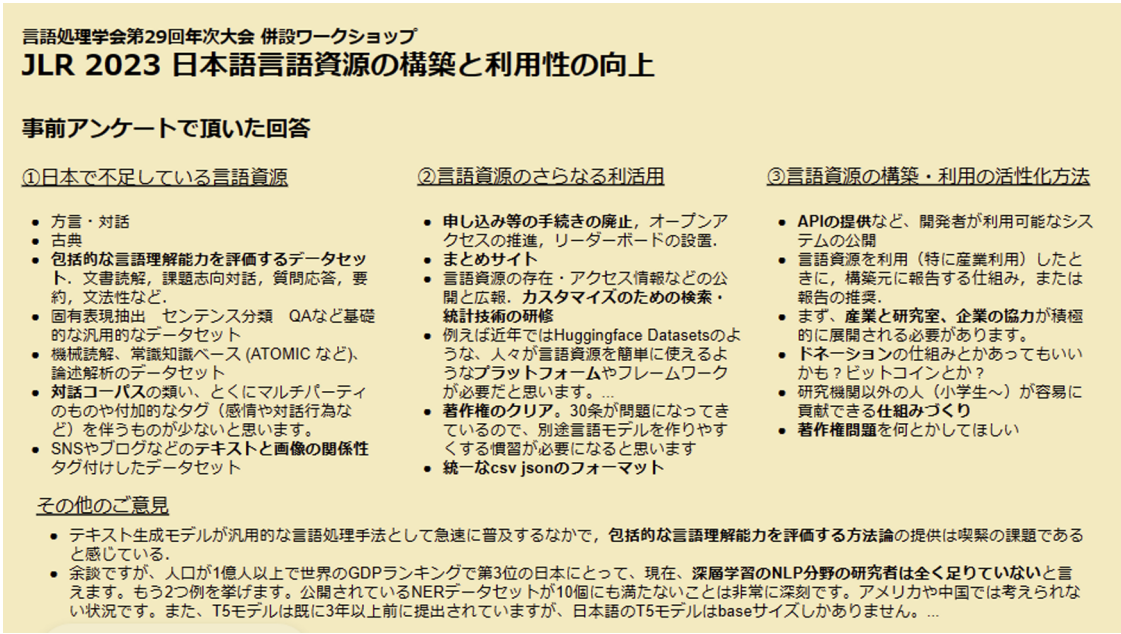

図 1　事前アンケートでいただいた回答

最後に，河原・久保・坂口・柴田・松田によりパネル（総合討論）を行った．総合討論においては事前募集質問についてディスカッションを行った．言語資源のさらなる利活用について議論がなされ，現代日本語を対象とした利用制限の少ないオープンライセンスなデータを構築するために何が必要かが議論された．また言語資源の構築・利用の活性化方法として，適切なデータ共有サービスや産学連携のあり方について議論された．さらに，事前学習モデル・大規模言語モデルを扱う上での実践的なノウハウの共有を進めていく必要があること，テキスト生成モデルの日本語の言語理解能力を包括的に評価する方法論とデータセットの整備が急務であること，などについて議論された．

3　おわりに

提案者一同，次回年次大会でも本ワークショップを開催するとともに，様々な企業の AI 技術勉強会と合同イベントを企画したいと考えている．今後も多くのご発表・ご支援を賜りたい．

謝　辞

JLR2023 の開催にあたり，言語処理学会第 29 回大会委員会の皆様には多大なご支援を賜りました．また，本イベントは株式会社 Studio Ousia・科学研究費補助金基盤研究 (A)「計算知と人知の融合による汎用言語理解基盤の構築」・国立国語研究所共同研究プロジェクト「実証的な理論・対照言語学の推進」との共催です．

参考文献

Oba, D., Yamada, I., Yoshinaga, N., and Toyoda, M. (2022). "Entity Embedding Completion for Wide-Coverage Entity Disambiguation." In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 6333–6344, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ri, R., Yamada, I., and Tsuruoka, Y. (2022). "mLUKE: The Power of Entity Representations in Multilingual Pretrained Language Models." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7316–7330, Dublin, Ireland. Association for Computational Linguistics.

Yamada, I., Asai, A., Shindo, H., Takeda, H., and Matsumoto, Y. (2020). "LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6442–6454, Online. Association for Computational Linguistics.

Yamada, I., Washio, K., Shindo, H., and Matsumoto, Y. (2022). "Global Entity Disambiguation with BERT." In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3264–3271, Seattle, United States. Association for Computational Linguistics.

略歴

浅原　正幸：国立国語研究所・東京外国語大学教授.

河原　大輔：早稲田大学教授.

久保　隆宏：アマゾンウェブサービスジャパン合同会社 Developer Relations Machine Learning.

坂口　慶祐：東北大学准教授.

柴田　知秀：ヤフー株式会社上席研究員.

松田　　寛：株式会社リクルート Megagon Labs, Research Scientist.