公開

研 究 終 了 報 告 書

「人間と AI の双方に扱いやすいことばの単位の創出」

研究期間: 2021 年 10 月～2024 年 3 月

研 究 者: 平岡 達也

1. 研究のねらい

人間が自然言語を分析するときには，自然文を単語などのまとまったことばの単位に分割すると取り扱いやすくなる．同様に，AI で自然言語を分析したり処理したりする場合にも，適切なことばの単位に分割を行うことで AI の処理性能が向上することが，近年の研究からわかっている．

一般的な自然言語処理では，人手で整備した辞書を用いて人間が理解しやすいことばの単位に自然文を分割してから AI システムの処理を行う．しかし，新語などの辞書に収録されていない未知語を正しく分割できないことや，データスパースネス問題によって，AI システムによる処理性能が低下するという課題がある．近年では，辞書を用いずことばの頻度の統計情報などを用いた分割方法も利用されている．これはデータスパースネス問題が解消できるなどの点で機械学習との相性が良く，AI システムの自然言語処理の性能向上に貢献するが，人間には理解が難しいことばの単位を使用する場合がある．

現在の AI 技術において人間が理解しやすいことばの単位と AI が扱いやすいことばの単位を両立することは難しい．一方で，AI 技術が国民に広く普及し，国民一人ひとりが AI システムの処理結果を解釈して判断する機会が増加している現代において，AI システムが人間には理解し難いことばの単位を用いることは上記のような判断ミスや，システムの使いづらさに直結するため，人間と AI の双方に扱いやすいことばの単位の解明と，それを用いた新たな AI 技術の普及が課題となっている．

本研究の目的は，AI が扱いやすいことばの単位を解明し，人間が扱いやすいことばの単位と折衷することで，人間による理解のしやすさと AI による自然言語処理能力を両立させるような新しいことばの単位を創出することである． 目的を達成するために，以下の 3 つの段階的な目標を設定し，研究に取り組んだ．

① AI が扱いやすいことばの単位を発見する機械学習の手法を開発する
② 人間と AI が扱いやすいことばの単位を比較し，その共通点を分析する
③ 人間と AI の双方が扱いやすいことばの単位を用いた自然言語処理のツールを開発する

2. 研究成果

(1)概要

本研究は，研究担当者の博士論文のための研究に起点をおいた取り組みである．段階的な目標設定のうち「①AI が扱いやすい言葉の単位を発見する機械学習の手法の開発」については，はじめに，自然言語処理において AI の性能が向上するようなことばの単位（専門的にはトークナイゼーションやトークン分割と呼ぶ）の最適化を行う手法を開発した．博士論文の研究自体については ACT-X の成果の範囲外であるが，本研究を拡張することで，ACT-X の活動期間の成果として新たなことばの単位の最適化手法の開発した．これらの研究により，AI に扱いやすいことばの単位が，性能を指標にした最適化という方法で獲得できるようになった．また，獲得された AI に扱いや

公開

すいことばの単位が，人間にとっても扱いやすいとは必ずしも言えないという観察が得られた．

この観察内容を量的に詳しく分析すべく，「②人間と AI が扱いやすいことばの単位を比較し，その共通点を分析する」項目を実施した．本項目では，複数のことばの単位に対する人間・AI 双方にとっての扱いやすさの違いに関するデータを実験によって収集した．実験結果より，人間は辞書に収録されているような，言語学的なことばの区切り方による単位が扱いやすいと自覚していることが分かった．一方で，実際に問題を解くときには AI に扱いやすいことばの区切り方も，人間にとって扱いやすいと判断されていることが分かった．特に，言語モデルと呼ばれるアーキテクチャを用いて作成されたことばの単位は，人間にとっても AI にとっても，問題を解くときに扱いやすいと判断されていることが確認された．

これらの分析を踏まえて，「③人間と AI の双方が扱いやすいことばの単位を用いた自然言語処理のツールを開発する」項目を実施中である．言語モデルを用いたトークン分割モデルをベースとして，人間にとっての扱いやすさを向上させるための工夫を取り入れた新たな手法を言語処理学会第 30 回年次大会にて発表予定である．また，ACT-X の直接的な成果ではないが，LLM 勉強会での大規模言語モデルの開発においても，本研究によって得られた成果を応用してトークン分割の手法の開発に貢献した．

さらに本研究課題を通して，AI がことばの表層的な情報の操作を非常に苦手としていることが分かった．例えば近年広く使用されている AI では，「COVID」ということばが 5 文字から構成されていることや，部分文字列に ID を含む，といった情報を認識することができない．本問題については現在検証結果を取りまとめており，言語処理学会第 30 回年次大会にて発表予定である．

（2）詳細

本研究での成果を，目的①②③に応じて以下の通り報告する．また④では，本研究を進める中で発見した問題を元に取り組んだ成果について報告する．

① AI が扱いやすいことばの単位を発見する機械学習の手法を開発する

本項目は，研究担当者の博士論文としての研究[1]を継続する形で取り組んだ．博士論文での研究では，文書分類や機械翻訳などの後段のタスクを解くためのモデル（後段モデル）の性能が向上するように，ことばの単位（トークン分割）を決定するような手法を開発した．従来の自然言語処理ではトークン分割の処理は前処理として取り扱われていたため，一度決定したトークン分割をあとから修正することはなかった（図 1 上）．この研究では従来の常識を覆し，解きたい問題に応じてあとからトークン分割を最適化する手法を提案した（図 1 下）．博士論文のための研究の大部分は ACT-X の期間以前の成果によるものであるが，開発した手法の振る舞いの分析や，実験によって獲得された AI に扱いやすいことばの単位の分析は ACT-X の活動に含まれる．

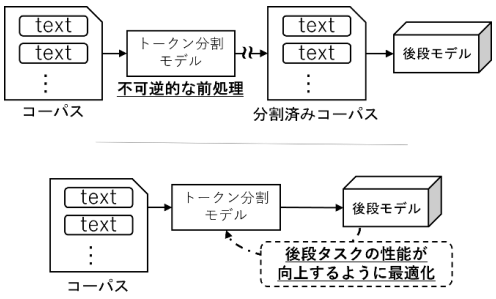

図 1　ことばの単位（トークン分割）の最適化の概要

博士論文の研究で開発した手法は，トークン分割のモデルとして，ユニグラム言語モデルと呼ばれるアーキテクチャを使用する必要がある．ユニグラム言語モデルは単純なアーキテクチャであるため，複雑な文脈などを考慮したトークン分割の決定はできない．ACT-X での研究においては，

公開

AIに扱いやすいことばの単位を獲得する必要があり，より表現力の高いアーキテクチャによるトークン分割のモデルを用いた，ことばの単位の最適化手法が必要となる．そこで，ACT-X の期間内では，使用できるトークン分割のモデルのアーキテクチャに制限がない，新たなことばの単位の獲得手法を提案した[2]．この研究では，従来手法がもつアーキテクチャの制限を取り払うための工夫を提案しただけではなく，AI に扱いやすいことばの単位の発見には多くの改善の余地が残されていることも示した．例えば，本手法を用いた場合の日本語の感情分類の性能（F1 値）は 86.36%であるが，仮に最も性能が高くなるようなことばの単位を選べていたとしたら，94.57%まで値が向上するはずであるという結果が得られている．

[1] Tatsuya Hiraoka. Task-Oriented Word Segmentation. Doctoral Dissertation, Tokyo Institute of Technology. 2022.
[2] Tatsuya Hiraoka. Tomoya Iwakura. Downstream Task-Oriented Neural Tokenizer Optimization with Vocabulary Restriction as Post Processing. arXiv, 2023. (Under Review)

② 人間と AI が扱いやすいことばの単位を比較し，その共通点を分析する

本項目では，AI に扱いやすいことばの単位が人間にとっても扱いやすいのかを検証し，共通点や相違点について分析する．AI にとって扱いやすいことばの単位は，項目①の手法を用いたり，複数のことばの単位を用いて AI を学習し，タスクでの性能を評価したりすることで抽出することができる．対して，人間にとって扱いやすいことばの単位については，実際にクラウドソーシングを用いてアノテーション作業を行い，収集する必要がある．

本項目の目的を達成するために，複数のことばの単位について (A)ことばの単位の適切性と(B)ことばの単位を用いたときの常識クイズの正答率・回答時間の収集のためにアノテーション作業を行った．データセットには，日本語の常識クイズが収録されている JCommonsenseQA を使用し，データセットに含まれる問題の全件（10,058 件）に対してそれぞれアノテーションを行った．本アノテーションでは，ことばの単位として辞書ベースで決定された単位（MeCab），頻度ベースで決定された単位（BPE, Unigram, MaxMatch），ある AI のモデルに最適化した単位（OpTok：上述の博士論文で開発した手法），ランダムに決定した単位の 6 件を採用した．

図 3 ランキングアノテーションの UI

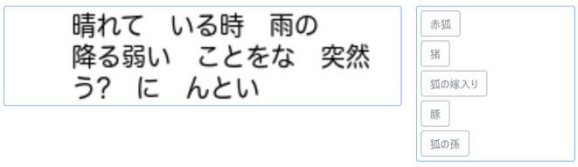

図 2 常識クイズの回答アノテーションの UI

(A)ことばの単位の適切性の収集には，図 2 のようなアノテーションツールを用意し使用した．6 つの異なることばの単位をランダムな順に表示し，アノテーターは区切りが適切だと思う順番に並

び替えるように指示される. (B)ことばの単位を用いたときの常識クイズの正答率・回答時間の収集には, 図 3 のようなアノテーションツールを用意して使用した. ことばの単位の区切りをアンダースコア記号で表示する(A)とは異なり, (B)では語順がわからないようにシャッフルして提示した. これは, アンダースコアで区切るだけでは, 容易にわかりやすいことばの単位に読み替えることができ, ことばの単位の違いによる結果の差が得られないためである. また, 扱いやすいことばの単位を用いている場合は, 語順の復元も容易であるという考えに基づいている.

本アノテーション作業の結果の一部を下表にまとめた. AI の解答性能, 人間の適切性・解答性能の値を比較することにより, AI は頻度ベースで作成したことばの単位が扱いやすく, 人間は辞書ベースで作成したことばの単位が扱いやすいことが分かった. ここから, 人間と AI のそれぞれに扱いやすいことばの単位が必ずしも一致しないことを定量的に示すことができた. 一方で, 人間の回答時間を見ると, 頻度ベースや AI モデルの最適化によって獲得したことばの単位が短い時間で回答できている事がわかる. ここから, 理解の速さという点に着眼した場合は, 人間と AI にとって扱いやすいことばの単位が一致する可能性が示唆された. またこの結果は, 頻度ベース(特に言語モデルを使用したもの)で作成したことばの単位が人間と AI の双方に扱いやすいことばの単位の折衷として相応しいことを示唆する. 本研究の成果はプレプリントとして公開しており[3], 現在国際学会の査読待ちである.

	AI:解答性能 [平均 F1 値]	人間:適切性 (A) [平均順位]	人間:解答性能 (B) [平均 F1 値]	人間: 回答時間 (B) [平均(秒)]
MeCab	42.48%	**1.86 位**	**94.30%**	6.07 秒
Unigram	**43.97%**	2.65 位	94.25%	**5.99 秒**
OpTok	43.38%	2.55 位	94.06%	**5.99 秒**
Random	41.94%	5.61 位	93.67%	6.93 秒

[3] Tatsuya Hiraoka. Tomoya Iwakura. Tokenization Preference for Human and Machine-Learning Model: An Annotation Study. arXiv, 2023. (Under Review)

③ 人間と AI の双方が扱いやすいことばの単位を用いた自然言語処理のツールを開発する

本項目では, ①及び②で得られた成果をもとに, ことばの単位を決定する新たなトークン分割ツールを作成する. ②の結果より, 人間と AI の双方に扱いやすいことばの単位が得られる分割方法の折衷として, 言語モデルを用いた方法の有効性が定量的に示された. そこで, 言語モデルを用いたトークン分割のモデルをベースとして, 人間の理解しやすさや AI にとっての扱いやすさと関係があると考えられる指標を新たに加えた手法を開発中である.

④ 本研究から発展した取り組み

AI が扱うことばの単位に焦点を当てた本研究では, 上記のメインストリームに加えて, AI の性能向上を目指したツールの開発にも取り組んだ. 自然言語処理の AI を学習する際には, 様々なことばの単位を用いることで, 性能の向上が得られることが知られている. これを実現するテクニックの一つとしてサブワード正則化という手法があり, 本研究では最長一致アルゴリズムによるトー

クン分割処理に対して，サブワード正則化を適用するための新たな手法を提案した．本研究は，自然言語処理の国際学会 COLING に採択済みである[4]．また本研究は，①で実施した AI にとって扱いやすいことばの単位の獲得するための新たな手法[2]を実現するためにも使用している．

本研究課題における調査を通して，近年自然言語処理で使用されている AI のモデルが，単語の表層的な情報を獲得できていないという問題を認知した．例えば，「COVID」という単語が 5 文字であったり，「ID」という部分文字列が会ったり，3 文字目が「V」であったりといった情報は，単語レベルのベクトル表現に適切に埋め込まれていない．この問題の詳しい検証を報告書作成時点で行っており，その成果は言語処理学会第 30 回年次大会にて発表予定である．さらに研究を発展させ，国際学会への発表も狙う．

ACT-X とは直接関係のない間接的な成果として，日本における大規模言語モデル（LLM）開発におけるトークン分割モデルの作成への貢献が挙げられる．分かち書きを行わない日本語においては，LLM が処理することばの単位を決定するトークン分割モデルの作成が，性能に大きな影響を与えると考えられる．研究担当者は，LLM 勉強会での LLM 開発[5]などに参画し，トークン分割モデルが人間にも理解しやすいことばの単位を持つように調整するべく，学習データや前処理・後処理の設計を行った．また，日本女子大学の研究グループが取り組むプログラミングコードを処理するための LLM 開発においても，トークン分割モデルの作成について議論し，国際会議 PACLIC への論文採択へと繋がった[6]．

[4] Tatsuya Hiraoka. MaxMatch-Dropout: Subword Regularization for WordPiece. COLING. 2022.
[5] 130 億パラメータの大規模言語モデル「LLM-jp-13B」を構築　～NII 主宰 LLM 勉強会（LLM-jp）の初期の成果をアカデミアや産業界の研究開発に資するために公開～．国立情報学研究所プレスリリース．2023 年．
[6] Teruno Kajiura, Shiho Takano, Tatsuya Hiraoka, and Kimio Kuramitsu. Vocabulary Replacement in SentencePiece for Domain Adaptation. PACLIC, 2023.

3．　今後の展開

本研究では，日本語の常識クイズのデータセットに対して，複数のことばの単位を用いた記述を用意し，適切性と回答内容・回答時間のアノテーションを行った．アノテーション済みの本データセットは，研究期間終了後に GitHub 上で全件を公開予定である．データセットを公開し，多くの研究者が自由に触れられる環境に置くことで，多面的な分析が得られると期待される．ACT-X 期間中の研究担当者による単独の分析だけでは発見できていない，新たな観点からの分析を得ることができれば，本研究は更に前進すると考えられる．

本研究では，目的③として分析結果に基づくツール開発を行っている．本ツールは 2024 年度中の公開を目指して整備を行うが，これはあくまでも分析内容の一部を用いた成果である．ACT-X の研究期間終了後も，外部の研究者を巻き込んだ上で収集したアノテーション済みのデータセットを引き続き分析し，新しい発見をもとにしたことばの単位の調査やツールの開発に活用していく．

本データセットは，ことばの単位の違いに対する人間の振る舞いの差を日本語で収集した初めてのデータセットである．本データセットを用いた分析を継続することで，自然言語処理分野における日本語の存在価値を高め，日本語での分析やツール開発の促進に繋げる．

公開

4. 自己評価

本研究の 3 点のねらいについて，以下の通り自己評価を報告する.

① AI が扱いやすいことばの単位を発見する機械学習の手法を開発する

本項目については，当初の予定通り博士論文として初期の手法を完成させており，達成済みである. また，初期の手法を拡張した発展的な研究も手掛けており，研究開始当初の見込みを上回る成果が得られている. 本研究対象は，AI のさらなる性能向上に向けて改良する余地が大いにあり，今後も ACT-X の研究で得られた成果をベースとして取り組み続ける.

② 人間と AI が扱いやすいことばの単位を比較し，その共通点を分析する

本項目については，複数のことばの単位を対象として，人間による扱いやすさのアノテーションデータの収集を行った. このアノテーション済みのデータセットは，本研究課題における最大の成果であると言える. アノテーションデータの統計的な分析については当初の予定通り実施済みであるが，言語学的な観点からの分析については十分な実施ができていない. これは，AI に扱いやすいことばの単位と，人間の回答時間が短いことばの単位に共通部分が見られるという，想定外の結果に対して分析の時間を充てたためである. 言語学的な分析については，研究担当者が単独で実施することは難しいため，今後も外部の研究者と連携し，引き続き取り組む予定である.

③ 人間と AI の双方が扱いやすいことばの単位を用いた自然言語処理のツールを開発する

本項目については，②で得られた分析結果の一部を用いて，新たなツールを開発し評価する段階まで成果が得られた. ツールの公開は ACT-X の研究期間終了後になる見込みである. 本ツールは，分析結果の一部のみを用いているため，当初の想定と見比べると部分的な達成にとどまる. これは，②でのデータ収集と分析に想定よりも時間がかかり，最終年度の開発期間へとずれ込んだためである. 収集したデータは研究期間終了後も分析のために活用を続け，得られた結果をもとに新たなツールの開発へと繋げていく予定である.

ACT-X を通した研究者ネットワークの構築という観点については，当初期待していたよりも多くの成果が得られた. 領域会議での異分野の研究者との議論から，新たな研究のアイディアが生まれたり，共同研究の繋がりが生まれたりしており，ACT-X の研究期間終了後も続くネットワークの構築ができている. また，ACT-X の研究を通してことばの単位に関する研究者として認知され，複数の大規模言語モデルに関する研究プロジェクトや，ことばの単位に関する共同研究への参画にも繋がっている. 報告書提出時点では，論文出版に至っていない研究も多いが，引き続き論文化や特許化を目指して取り組みを続ける予定である.

5. 主な研究成果リスト

（1）代表的な論文（原著論文）発表

研究期間累積件数： 5件

（2023年度末までの投稿予定論文を含めると7件）

代表的な論文3件

1. Tatsuya Hiraoka. MaxMatch-Dropout: Subword Regularization for WordPiece. In

公開

Proceedings of the 29th International Conference on Computational Linguistics (COLING), pages 4864–4872, Gyeongju, Republic of Korea, October 2022.
自然言語処理において，複数のことばの単位を用いてモデルの学習を行うことで性能が向上することが知られている．このテクニックはサブワード正則化と呼ばれており，本論文では最長一致法に基づくトークン分割器で利用可能なサブワード正則化の手法を提案した．実験より，文書分類や機械翻訳などの自然言語処理のタスクで性能の向上に貢献することが示された．
2. Tatsuya Hiraoka, Tomoya Iwakura. Tokenization Tractability for Human and Machine Learning Model: An Annotation Study. arXiv:2304.10813. 2023.
AI の性能が向上するようなことばの単位が，人間にとっても扱いやすいものであるかを調査し分析した．日本語の常識クイズに関するデータセットについて，複数のことばの単位を用いて AI を学習したときの性能の比較と，人間による適切性のアノテーション・正答率・回答時間を収集することで，比較を行った．結果より，人間と AI がそれぞれ扱いやすいと感じることばの単位に違いがあることを定量的に示した．また，回答時間という観点から，一部人間と AI に共通する嗜好があることが分かった．
3. Tatsuya Hiraoka, Naoaki Okazaki. Knowledge of Pretrained Language Models on Surface Information of Tokens. arXiv:2402.09808. 2024.
AI がテキストを処理するうえで，単語の長さや，何文字目にどの文字があるかといった表層的な情報を適切に処理できているかを調べた．広く用いられる大規模言語モデルの多くは，表層的な知識を十分に持っておらず，自然言語処理タスクでの性能低下の原因になりうることを指摘した．こうした問題は，トークン分割の観点から解決できる可能性がある．

（2）特許出願

研究期間全出願件数： 2 件（特許公開前のため詳細なし）

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

- 第 10 回 AAMT 長尾賞学生奨励賞. Tatsuya Hiraoka. Task-Oriented Word Segmentation.