第8回 国語分科会言語資源小委員会（Web開催）議事録

令　和　7 年　8 月　4 日
10 時 00 分 ～ 12 時 00 分
文部科学省3階 3F2特別会議室

〔出席者〕

（委員）相澤主査、石川副主査、植木、神永、武田、前川、森各委員（計7名）
（ゲスト）国立国語研究所（山崎客員教授、小木曽教授）
（文部科学省・文化庁）山田国語課長、中田課長補佐、武田主任国語調査官、
鈴木国語調査官、町田国語調査官ほか関係官
※ 相澤主査、石川副主査、神永、武田、森各委員及び事務局は、
3F2特別会議室にて参加。

〔配布資料〕

1 第7回国語分科会言語資源小委員会議事録（案）（委員限り）
2 BCCWJ2の設計と構築（前川委員御提出資料）
3 大規模言語モデルと言語資源（相澤主査御提出資料）

〔参考資料〕

1 言語資源小委員会における主な審議事項及び当面の予定（案）
2 今後における日本語のデジタル言語資源の整備・活用の在り方（報告）
（令和7年3月17日 文化審議会国語分科会）

〔経過概要〕

1 事務局から定足数の確認があった。
2 事務局から配布資料の確認が行われた。
3 前回の議事録（案）が確認された。
4 前川委員から配布資料2「BCCWJ2の設計と構築」に基づいて説明があり、質疑応答が行われた。

5 相澤主査から配布資料3「大規模言語モデルと言語資源」に基づいて説明があり、質疑応答が行われた。

6 二つの説明を踏まえ、BCCWJの拡充や大規模言語モデル、生成AIをめぐって意見交換が行われた。

7 次回の言語資源小委員会について、令和7年9月9日（火）15時から17時まで、対面及びオンラインで開催する予定であること、今期後半の御予定伺いを発送する予定であることが確認された。

8 質疑応答及び意見交換における各委員の発言等は次のとおりである。

○相澤主査

定刻になりましたので、ただ今から、第8回、今期2回目の文化審議会国語分科会言語資源小委員会を開催いたします。

本日も、文部科学省の会議室に実際にお越しくださった方とオンラインで御参加くださった方とがいらっしゃいます。また、傍聴者の方々もオンラインでこの会議を御覧になっていることを御承知おきください。

さて、本日、お手元の議事次第にございますとおり、言語資源の整備と活用、有識者による事例等紹介、その他という内容で協議を行いたいと考えております。

事前に資料を御覧になられた方もいらっしゃるかと思いますが、本日は、この小委員会の審議のきっかけとなりました、国立国語研究所の現代日本語書き言葉均衡コーパスの拡充について、その設計と進捗状況の報告を前川委員より頂く予定でございます。本日は、その実際のコーパス拡充作業に携わっていらっしゃいます山崎先生と小木曽先生にも御参加いただいております。

二つ目の報告としましては、私、相澤から、「大規模言語モデルと言語資源」と題して、大規模言語モデルの開発の近況についてお話をさせていただきます。

早速ですが、議事に入ります。まず、本日は前川委員から、現代日本語書き言葉均衡コーパスの拡充について、その設計と進捗状況をお話しいただくこととなっております。配布資料2の形で事前に配布しておりますので、よろしくお願いいたします。

○前川委員

おはようございます。国語研究所の前川です。「BCCWJ2の設計と構築」－こういうタイトルでお話をさせていただきます。私が一応組織の長として筆頭になっておりますが、実際の仕事を担当しておりますのは小木曽と山崎、そしてSNSの部分に関して落合が関係しておりますので、4名の連記と、連名の発表とさせていただいております。

それでは、早速始めたいと思います。今日のお話の構成ですけれども、こういう形で話そうと思います。前半、時間にして3分の1程度を、ちょっと比較する必要がありますので、BCCWJの話－昔のコーパスの話をさせていただいて、その後、後半、3分の2ぐらいの時間を使って、現在進行中のBCCWJ2の話をさせていただこうと思っております。

最初に、これが全体の比較を1枚にまとめたようなスライドです。御存じの方も多いと思いますが、BCCWJというのは、2006年から2010年に掛けて科学研究費特定領域研究というので構築したコーパスです。いろいろなレジスターの日本語を取ってきて、全体で1億語規模のもので、公開を前提としたものとしては我が国で初の均衡コーパスであるということです。幸い、公開後、よく利用されておりまして、この前ちょっと調べてみましたら、Google Scholarで四千何件ぐらいの論文が引用してくれているようです。一方、BCCWJ2、今日のお話の主眼ですが、そちらは文化庁国語課さんからの委託事業である「信頼できる言語資源としての現代日本語の保存・活用のためのデジタル基盤整備事業」というもので、それを資金として現在構築を進めているものです。こちらは、この後詳しくお話しいたしますが、主に書籍を中心として1億語規模、要するにBCCWJの規模を倍にするということです。それから、さらにプラスアルファの部分も出てまいります。こちらはBCCWJの時の経験を生かして設計と構築の両面を合理化して進めているということでございます。

早速、その古い方のBCCWJ－場合によってBCCWJ1と言うこともありますが、それについての説明をさせていただこうと思います。

これについて、先ほど申しましたように、実はボトムアップの開発計画であったということです。当時、国語研では話し言葉のコーパスであるCSJというのの構築が終わったところで、それが比較的うまくいって、世の中にも広く受け入れられたので、今度は書き言葉のことをきちんとやろうというようなことで始まりました。ただ、当時、国語研の中で必ずしも言語資源というものに対して主要な事業として認知されていたわけではなくて、「外部資金でやるならどうぞ」みたいなことを言われながらやったというような面もございます。

何分、初めての仕事ですので、非常にいろいろな面で試行錯誤を行いました。むしろ試行錯誤を積極的に行おうと考えて、いろいろな問題に関して検討を進めたということでございます。その下に箇条書でいろいろなことが書いてございますが、全部触れると時間が長くなりますので、こういったいろいろな問題を取り扱ったということです。今日、後半の話で関係してくるところは、やはり最初のバランスということをどう考えるのかという話、それから、最後に出ております権利処理をどうするかというようなこと、ここら辺は実際問題として大変いろいろと複雑なことが生じてきたということでございます。

実際のBCCWJの構成というものを示しているのが、この図でございます。全体1億語が三つのサブコーパスに分かれておりまして、一つは出版サブコーパス、これは出版データからランダムサンプリングを行って取ってくるもので、3,500万語ぐらいです。それから、もう一つは図書館－これは東京都下にあります全ての公立図書館の所蔵している図書を、1986年から2005年の間のものを母集団としてランダムサンプリングを行うというものです。そのほかにいろいろ、日本語を研究する目的上必要なものを集めたのが特定目的サブコーパスというもので、こちらが3,500万。合計で1億語ということになります。

そこからいろいろなサンプルを取るわけですけれども、その中で、これは余り重要なことではありませんが、この後話が出てきますので、2種類のサンプルを取っています。長さについて2種類のもの－固定長サンプルというものと可変長サンプルというものを取りました。固定長は1,000字に固定しておりまして、可変長は、文章の構造を反映した節だとか章だとか、かなり長い範囲を取っています。上限は1万字としております。固定長の方は、要するに、例えば漢字語でも何でもいいんですけれども、母集団でどれぐらい存在するのか、あるいは使用率がどれぐらいなのかということをきちんと推定するためにこういったものを取ったということです。可変長はもう少し長く取っておりますので、いろいろな言語的な研究のために取ったということです。それから、先ほどのサブコーパスごとにそれぞれ、固定長と可変長両方取ったものとそうでないものとがございまして、それを右の表に示しております。

それから、取ってきたサンプルを形態素解析に、あるいは形態論情報付与というものを行います。要するに、単語に切る－切って品詞を与えるという作業ですが、その際にも、これは日本語の膠着語としての特性上、1種類の単語だけではうまくいきませんので、短単位と言われるものと長単位と言われる2種類の形態論情報を付与したということです。下に例が挙がっておりますが、例えば「日本語コーパスについて」というような部分は、短単

位で切りますと、下にありますような「日本」「語」「コーパス」「に」「つい」「て」というように切れますし、長単位で切りますと、「日本語コーパス」全体は1個の長い複合名詞としてまとまりますし、「について」、助詞の部分もいわゆる複合辞、複合助詞として一つにまとまるというようなものでございます。あと、サ変動詞、「解説する」みたいなものも、短単位では「解説」と「する」に切れますが、長単位では1個にまとまるというような、そういう違いがございます。これは実際、研究目的によってどちらを使うかということで研究者の方々は使い分けておられると思います。それから、最後に、一番下に書いてありますように、全体としては、当時、現在ほど解析精度が高くございませんでしたので、自動解析ですと、ジャンルによりますけど、95から98%ぐらいの精度しか出ませんので、そのうちのコアという部分を設定いたしまして、その部分に関しては人手で精度向上を目指しました。99%以上というのを目指しました。

それから、最後に権利処理になりますが、これが現代語コーパスでは一番大変な問題になっております。それで、当時もいろいろな社会情勢の変化があって、権利処理についてはいろいろな意見が対立している状況でした。ですので、専門家の方に伺っても、必ずしも一貫した意見が返ってこない－これは現在も同じであろうかと思います。BCCWJでは結局、誰にも文句を言われない、言い換えると、「ばか正直な処理」というものを目指しました。書籍を中心に2万4,000件を処理いたしました。右のグラフは、上の方が、これ、プロジェクトが終わる1年ぐらい前の段階ですけれども、2万4,000件のうち、どれぐらいの連絡先を探し出して、実際に連絡を取れたかというグラフでございます。下の方は、連絡済みのうち、許諾していただけたものが約7割で、拒否されたものが5%、それから返事が返ってこないのが残りというようなことを示したものでございます。最終的にはもうちょっと高い許諾率になったかと記憶しております。

これは開発体制を示しております。先ほど申しましたように、科学研究費のプロジェクトでしたので、データを作るだけではなくて、それを解析する、コーパスの評価というグループもございました。項目Aのコーパスの構築というところがコーパスの構築及びそれに関連するツールなどの開発を行う部分でして、そのうち、特にデータ班というところが国語研の中に設置されておりまして、その中がさらにサンプリングするグループ、形態論解析を行うグループ、著作権処理を行うグループ、それから電子ファイルを作るというグループの四つに分かれて活動を続けておりました。山崎さんは、このうちのデータ班の取りまとめを行ってもらったということです。常勤と非常勤を合わせて、当時は総勢25名ぐらいの体

制で、かなり大きなグループとして活動しておりました。

公開方式ですが、これは皆さん御存じのことかと思いますが、テキストと形態論情報を「少納言」、「中納言」という形で無償公開しております。それから、データ全体を有償でも公開しておりまして、これは契約によっては商業利用も可能という形になっております。それから、主に有償公開をしていた方を対象として、その後、国語研で開発いたしました様々なアノテーションとかメタ情報とかを追加で公開しております。これは国語研のレポジトリーで随時公開しております。

このグラフは「中納言」の利用実績でして、毎年、順調に伸びている。登録ユーザー数が多分、2023年までのデータですので、現在ですと6万人を突破していると思います。それから、年間の検索件数も300万件－300万回を昨年度は突破していたんじゃないかと思います。

ここまでがBCCWJで、この後、BCCWJ2の話に移りたいと思います。

開発の動機ですけれども、BCCWJを作ってもう20年たちますので、現代語コーパスとしては古くなっている部分がある。それから、経年的な調査ができない。規模的にも1億語というのは小さなものであるというようなことがございました。そういうところに文化庁国語課による概算要求をしていただきまして、先ほど紹介いたしました事業が始まったということでございます。事業の内訳といたしましては、2006年から2025年までのまた20年間分の日本語データを追加してサイズを倍にするということでございます。その前提といたしまして、前期の文化審議会国語分科会の中での報告書で、こういったことをきちんとしろという報告をしていただいたのも大きなモチベーションになったかと考えております。

そして、先ほども触れましたが、BCCWJ2というのは、1を前提とした拡張計画になっております。つまり、BCCWJ1の経験をいろいろな意味で生かす形で開発を進めております。ただ、一つ前回と非常に異なっておりますのは、前回BCCWJ1を作ったときは独立行政法人だったんですが、現在は大学共同利用機関法人に移管されておりまして、要するに、非常にアカデミックな性格が強くなっております。前のように、所内の人間を組織して、一貫して一つの仕事だけ集中的に行うというようなことが非常に困難な状況になっておりますので、それに即した開発体制を現在は取っているということ、これは後に触れることにいたします。

それで、ここからが重要なところになってまいりますが、BCCWJ2ではどういうサンプルで構成するかという問題でございます。先ほど簡単に紹介しましたように、BCCWJ1の

方にはいろいろなレジスターが入っている。細かく分けると、たしか25ぐらいのレジスターだったと思いますが、それよりはBCCWJ2は大分単純化されております。それは、これから説明いたしますように、主に社会的な情勢の変化を反映しているものです。まず、新聞ですけれども、これはもう新聞各社がそれぞれ自分たちのデータを事業化して、それを販売されておりますので、その意味ではもうデータが入手できないということで、今回は対象から外しております。それから、雑誌ですけれども、BCCWJ1のときには母集団があった－「雑誌新聞総かたろぐ」というものが刊行されていたんですが、2019年まででそれが廃止されてしまいましたので、同じような形でのサンプリングは不可能ということで、この二つ－新聞と雑誌はBCCWJ2の対象とはしない、あるいはできないということになりました。ただし、国語研が別途、小木曽を中心に開発しております昭和・平成書き言葉コーパス（SHC）というのがございまして、これは十分ではございませんが、新聞と雑誌を8年間隔でサンプリングするということをやっておりますので、そちらである程度補うことは可能かという形になっております。

それから、2点目といたしまして、漫画とかアニメの日本語、これは当然、いろいろな意味で、日本語の実態を知るという意味でも、また日本語を研究するという意味でも重要な、あるいは日本語教育の観点からも重要なものであると思いますが、現時点ではデータ入手のめどが立たないということ、また、入手したとしても、これは性格がかなり書き言葉とは異なりますので、それに関する別途、独自の設計が必要になるであろうというようなことがございまして、現時点では対象として考えておりません。

それでは、一体何がBCCWJ2に含まれるかということでございますが、基本的にはBCCWJ1、今下に出ておりますが、その左上にあります出版サブコーパスを継承して拡大するという形で進めております。つまり、出版データに基づいて、それをランダムサンプリングしてもう一度やるということでございます。それに加えまして、下の特定目的サブコーパスの中にございます教科書の部分も継承・拡大をいたします。そして、これが大きな違いになってまいりますが、SNSのデータというものを別途追加していきたいと考えております。1億語の対象になりますのは出版サブコーパスの部分になるわけですけれども、それに加えまして、SNSなどのデータを加えていきたいということでございます。詳しくはこの後、御説明いたします。

最初に、継承の部分に当たります書籍をどうやってサンプリングしているかということですが、基本的にはBCCWJ1のときと同じ方法でサンプリングを行います。ただし、この20

年間に出版が大きく変わりました。電子書籍というものが広く普及しております。ただ、電子書籍に関しましては母集団が設定できませんので、結局、前回と同じような出版データの存在する紙の書籍というものを対象にしております。母集団といたしましては、前回と同様、国会図書館が出しておりますJAPAN／MARCというものを母集団として利用させていただいております。今下に示されております図は、そのJAPAN／MARCから取ってまいりました各年の書籍出版の総数でございます。大体毎年10万とか12万ぐらいの出版が行われていて、COVID－19の頃にちょっと減ったというような感じでございます。右端が一番減っているのは、データが足りない－まだ完璧でないということです。

そのサンプリングのやり方ですけれども、2025年のデータというのがまだ存在しておりませんので、全体を一つの母集団とするのではなくて、1年ごとに母集団を想定して、それぞれ500万語を抽出して、合計20年で1億語という形のサンプリングを行います。それから、前回は総ページ数というものを推定して、それを単位にサンプリングをやっていたんですが、そういう面倒くさいことは今回やめまして、書籍を単位として抽出しております。もちろん書籍は長いものも短いものもございますので、必ずしも等確率での抽出になりませんけれども、余りそれが問題になることもないだろうということで、こういうサンプリングをやっております。それから、固定長と可変長は前回と同じく両方取りますが、取り方を若干簡素化しております。それから、書籍といってもいろいろございまして、幾つかのものを除外しております。そこに挙げておりますようなものを除外している。これは前回とほぼ同様です。そして、先ほどコアというものを御説明いたしましたが、今回もコアを設定いたしまして、全体で400万短単位程度をコアに設定する予定であります。

今出てまいりましたグラフは何かと申しますと、先ほどのJAPAN／MARCの母集団から実際にサンプリングをした。そして、その中のどういうものが入っているかということを日本十進分類で色分けしてお示ししているということです。BCCWJ1のときと大きくは変わっておりません。

それから、書籍データ以外の部分ですけれども、教科書データといたしましては、今お示ししているような形のものを加える予定でございます。BCCWJ1との違いといたしまして、BCCWJ1の時には、教科書データ全体もやはりサンプリングをいたしまして、ページ単位のサンプリングをしたんですけれども、今回は全文を入れるということになっております。それから、過去3回の学習指導要領改訂に対応した3年分の全教科を入れるということでございます。サイズとしては、1から3、BCCWJ1のものも含めて全体で2,000万語に

なる予定でございます。

それから、これが大きな違いになってまいりますが、SNSのデータでございます。SNSというのは、皆さんお気付きかと思いますが、BCCWJ1がちょうど終わった後に生まれたメディアなんです。BCCWJ1では2005年までのデータを取ったわけですけれども、その後、ツイッターが2006年、フェイスブックも2006年、LINEが2011年という形で、その後に生まれたメディアです。これも言うまでもありませんが、現代の言語使用の実態のある側面を非常にクリアに表しているデータであろうと思います。総務省が、どれぐらい国民がこういったものを使っているかという調査を行っておりまして、これはなかなか複雑な調査なんですけれども、簡単にまとめますと、年齢をごっちゃにして分析すると、国民の43％がツイッターもしくはXを使っている。それから、91％がLINEを使っているということで、実は私もLINEは使いますので、そんなものだろうなという気がいたします。いろいろなニュース番組とかで社会の動向とかを分析するのにこういったデータが使われているのは御存じのとおりです。ただし、メディアによりまして、データの入手可能性とか、あるいはそれに伴う価格設定は様々でございます。

今御覧になっていただいている図は、このSNSを共著者の1人の落合が、下に引用しております論文の中で分類している図でございます。詳しい説明は避けますが、あるテキストの受け手が人数において多いか少ないか、これが横軸です。それから、特定の人を念頭に置いているか、不特定を対象にしているかということを縦軸にして、各種のメディアを、SNSを分類したと、そういう図でございます。この中で今回対象といたしますのは、aとbと書いてある部分、この部分でございます。ここに関しては現実的にデータが入手できるということでございます。

実際、一つはLINE、aの部分です。この部分に関しましてLINEのデータを収集いたします。ただ、これ、自動収集は不可能ですので、協力者何十名かの方にお願いして、許諾を頂いた上で、3年計画でテキストチャットのデータを収集しております。現在の協力者としましては、東京出身の比較的若い人500名程度を対象としております。今後、対照群として関西方面でもデータを収集するかもしれないということでございます。サイズといたしましては、2024年から26年の3年間で70万語程度を収集しようと考えております。こういう方法ですので、あまり大量には集められないということです。

もう一つ、bのところです。ここの部分ですね。ちょうど座標の真ん中ぐらいに当たるところですけれども、これに関しましては、本来X（ツイッター）が占めている部分ですけれども、

ツイッターは本文の再配布を許してくれませんし、それから、APIを利用しようと思うと、かなり高額の利用料を請求されますので、その代わりとしてBlueskyというものとMisskeyのデータを収集中です。Blueskyは世界的なもので、Xの代替メディアとして広く利用されております。Misskeyは国内のものです。これらのものから、これはAPIを通じてデータを収集することができますので、こちらはかなり多くのもの、1年間でそれぞれ50億語、15億語というような大きなサイズのデータを取る予定でございます。

それで、次にアノテーションあるいはメタデータについて簡単に説明いたします。形態論情報に関しましては、短単位と長単位の二重解析というのはBCCWJ1の時と同一でございます。短単位は従来どおりMeCab＋UniDicで解析を行いますが、この20年間で辞書が飛躍的に充実しておりますので、自動解析の精度は多分飛躍的に向上している－本当に飛躍的に向上していると言って問題ないと思います。それから、長単位に関しましては、前回はいろいろな手段で解析したんですけれども、今回は専用の解析器を新規に開発して利用する予定でおります。それから、コアについては、先ほど申しましたとおり、400万短単位程度のサイズを設定する予定でございます。それから、タグですけれども、これは余り詳しくは先ほど説明いたしませんでしたが、BCCWJ1のときには非常にいろいろなタグとかメタデータを準備いたしました。いろいろな目的で使われるだろうという想定下で、これもあったらいい、あれもあったらいいということで様々なものを作成いたしましたが、実際には余り使われなかったものが多いということで、今回は基本的なものに集中して効率的に構築を進めたい。特に「中納言」での検索に用いるデータを優先して付与するという形で進めております。

最後に、一番大きな問題で、権利処理ですけれども、BCCWJ1の時代は大変厳しい時代でございました。先ほど申しましたように、非常に高い権利処理コストを負担して、ばか正直な処理をしたということでございます。その後、BCCWJのプロジェクトが終わった後に著作権法が2回改正されております。その改正の方向といたしましては、全体として情報処理での研究とか産業化を念頭に置いて権利制限の規定を拡大する、つまり、使う側が使いやすくするという方向での改定が行われております。

その新たに導入された権利制限規定のうち、我々に深く関係するものを三つここに挙げております。一つは、2012年に改正されました30条の4項というのがございまして、思想又は感情の享受を目的としない利用の場合、権利制限が行われると、つまり、自由に利用できるということでございます。それから、47条の4、これは余り直接の効果はございませ

んが、公衆配信のために計算機の中にキャッシュのような形で情報蓄積をするというような場合、これはいいということでございます。それから、47条の5、これはいわゆる軽微利用というもので、電子計算機による情報処理その他の結果に付随するもの、軽微な利用というものは許可を取らなくてもいいということになっております。こういったものを積極的に利用してBCCWJ2の権利処理を行うということでございます。これらの規定によって許される範囲内で構築して配信するということを原則としております。これによって開発コストを軽減して、またコーパスの価値も高めたいと考えております。利用価値が高いものにしたいということを考えております。

実際にこういう権利制限の規定の範囲内でどういう形で公開をするか、利用していただくかということですけれども、まず、一応、全体の電子媒体での有償配布というのは行います。これに関しましては、申込みをしていただくユーザーの方と契約を結びまして、30条の4項、つまり、中身を読むんじゃなくて、思想又は感情の享受を目的としない利用をいたしますということを条件に利用を許諾するという形を考えております。それから、一方、これが大多数になりますけれども、「中納言」等による無償でのオンライン公開に関しましては、表示される文脈をBCCWJ1―これ、現在、前後500語まで見ることが可能なんですけれども－最大500語ですね。それをぐっと制限することによって軽微利用の範囲に収め、そのことによって著作権処理を不要化しようと考えております。ちなみに、先ほどもちょっと触れました昭和・平成書き言葉コーパス（SHC）の公開方式がこれでございまして、ここでは確か前後30短単位という範囲での公開をしております。

それで、こういう全体の方針は今御説明申し上げたとおりですけれども、実際にBCCWJ1を現在使っているユーザーはどれぐらいの文脈長で検索しているのかということをちょっと調べてみました。「中納言」の検索ログ、278万件ほどございますが、それを分析した結果が右の、今の表のとおりで、20というところにほとんど集中しているんですが、この前後20短単位というのは「中納言」のデフォルトの使用です。皆さんほとんどデフォルトで使用されて、不便は感じておられないのかなという感じでございます。一部に50とか100を設定して検索される方もおられます。

では、累積で見るとどうかということですけれども、横軸が前後の文脈の範囲です。全部の検索の何％ぐらいがそれでカバーされたかということを表したグラフを、短単位と長単位と文字列検索という形で示しております。御覧いただきますと、例えば前後50語－50短単位ですけれども、短単位検索で95％、長単位検索で86％、そして文字列検索で

あれば87％程度の需要をカバーするということが分かっておりますので、これぐらいの範囲で公開すれば、著作権処理が、権利処理が不要になるのではないかと考えております。これ以上の文脈を必要とするユーザーはもちろんおられるわけですけれども、談話分析とかやりたい方はおられるはずですから、そういう方には有償版を先ほどの制約の範囲で、制約を行った上で利用していただくということを考えております。

最後に、これがBCCWJ2の今後のロードマップの案でございます。現在、2025年の途中でございます。比較的順調に進んでおりまして、今年度末には最初の部分的なデータ公開を予定しております。以後、ここに書いてあるような形で進めていって、最終的に2028年度で構築を完了するという予定で進捗しているところです。

これがまとめでございます。一々読みませんが、最後に書いてありますように、今後、26年度からになるかと思いますが、成果発表会的なイベントも開催して広報に努めたいと考えております。

以上、大体30分かと思います。以上で、私の話はここで終わりにさせていただきます。

○相澤主査

前川委員、ありがとうございました。

では、ただ今の御説明につきまして、御質問等ありましたら、よろしくお願いいたします。御意見、御感想を伺う時間は後ほど取る予定でございますが、記憶が鮮明なうちにということがありましたら、是非お願いいたします。いかがでしょうか。

○神永委員

少しよろしいでしょうか。大変貴重な御報告、ありがとうございます。辞書編集者としましても非常に興味深い内容ですので、是非完成していただきたいなと思っているんですけれども、まず1点なんですが、BCCWJ1の方なんですけれども、精度が99％で解析できるというようなお話だったと思うんですが、これは現在も何かその精度を上げるための作業を続けていらっしゃるんでしょうか。

○前川委員

99％というのはコアの部分の目標値で、一応そこに達しているということでございます。それで、全体としてはもうちょっと低いということでございます。それで、BCCWJ2の公開に

合わせまして、BCCWJ1の方も形態素情報を再分析いたしまして、アップデートするという計画が現在進んでおります。

詳しくお知りになりたければ、小木曽か山崎の方から補足説明いたします。

○神永委員

ありがとうございます。それはまた別の機会にお教えいただければと思うんですけれども、もう一点なんですが、ちょっと私、聞き逃してしまったのかもしれませんけれども、このBCCWJ1とBCCWJ2なんですけれども、追加というようなことをおっしゃっていますが、2025年から逐次公開していくということは、BCCWJ2で集めたものをBCCWJ1の方に少しずつ追加しながら、最終的に両方合わせた合体したものが2028年に完成すると言いますか、公開される－そういう感じなんでしょうか。

○前川委員

そこは我々もいろいろ考えたところなんですが、いきなり混ぜるということはしないつもりです。別のコーパスとして公開して使っていただくという形で。

○神永委員

そうなんですね。

○前川委員

最終的にも、多分、山崎さん、小木曽さんたちの考えでは、別コーパスとして、BCCWJ2という一つのコーパスとしてまとまった形で公開すると。ただし、私どもの「中納言」なんかでは串刺し検索ができますので、まとめて検索するということももちろん可能です。

○神永委員

ありがとうございます。その串刺しができるというのは非常に有り難いことです。

○前川委員

現在でも串刺しはできますので、いろいろ、例えば歴史コーパスなんかとですね。

○神永委員

はい。ありがとうございました。

○相澤主査

ありがとうございます。

ほかはいかがでしょうか。よろしいですか。

○石川副主査

石川です。ありがとうございました。幾つか伺わせてください。

まず1点目として、現在のBCCWJは13レジスター構成になっているわけですが、BCCWJ2は、書籍、教科書、そしてSNSなどの打ち言葉、これら3レジスターになるという理解で正しいでしょうか。

○前川委員

はい、そのとおりです。

○石川副主査

ありがとうございます。

2点目として、書籍に関して、現在のBCCWJは、書籍データの収集に当たり、「生産」と「流通」の観点を分け、さらにはそこにベストセラーも含めて多面的に捉えているかと思うのですが、BCCWJ2の書籍データの収集は「生産」面に限るという理解でよろしいでしょうか。

○前川委員

そういうことになるかと思います。

○石川副主査

ありがとうございます。

3点目です。BCCWJ2には大いに期待をいたしているところですが、これまでの御説明を踏まえますと、現在のBCCWJとはかなり内容的が変わるように思います。そうした場合に、出来上がったものをBCCWJ2と呼ぶのがよいのか、あるいは少し名前を変えたもの

にする方がよいのか、国語研ではどのようにお考えでしょうか。

○前川委員

その内容が実際同じでないわけですよね、それで2と付けた。

○石川副主査

なるほど。

○前川委員

先ほどの御質問と同じですけれども、一つにしないというのは、そういう考えの表れと御理解いただければと思います。

○石川副主査

分かりました。

4点目ですが、BCCWJ2の文脈表示について「前後50語」と伺いましたが、これはデフォルトで表示されるのは50語だがユーザーが更に広げることも可能なのか、あるいはマックス（最大値）が50語になるという意味なのか、どちらだったでしょうか。

○前川委員

これはマックスです。

○石川副主査

マックス50と。

○前川委員

はい。BCCWJ1は著作権処理が済んでおりますので、存在するサンプルは全部、やろうと思えば表示できるということなんですけれども、今回はそれをいたしませんので、デフォルト50＝マックス50とするか、あるいはデフォルト20・マックス50とするか、それは分かりませんけれども、ともかくマックスです。

○石川副主査

ありがとうございます。

○前川委員

それから、ちなみに、50もちょっと、今の時点で決まっているわけではなくて、そこら辺にするかなというような検討を進めているという段階です。

○石川副主査

ありがとうございます。

5点目は、著作権処理の作業を合理化するという御方針についてです。御指摘のように、コーパスに収録する著作物の著作権処理に関しましては、従来、決まっていないことが多いがゆえに、作る側が過剰に心配して、本来は必要でない手続まで踏んでいた部分があろうかと思います。その意味で、BCCWJ2の新しい作業方針は納得の行くものと存じますが、この点については、法律の専門家の皆さんの見解も一致していると理解してよろしいでしょうか。

○前川委員

これも意見が収れんしていない。

○石川副主査

なるほど。

○前川委員

人によって結構違うということです。私がBCCWJ1の頃から一番個人的に親しく、友人でそういう人間がおりまして、その男は「2回の法改正で、前川さん、これ完全に大丈夫ですよ」と言ってくれるんですが、世の中には「いや、そうではありません」という、両方の解釈が今もあります。

○石川副主査

分かりました。新しい法律の趣旨を踏まえてということでよく理解いたしました。

最後の質問です。BCCWJ1の開発時には、「項目B」として、「コーパス日本語学の創生」が掲げられ、様々な研究班が作られ、日本語の学界自体が非常に盛り上がったことを懐かしく思い出すのですが、今回のプロジェクトでもこうした研究プロジェクトを立てていく御計画はおありなのでしょうか。

○前川委員

いえ。今回は本当、構築に関する委託事業ですので、今回は完全に構築に集中いたします。ただ、今後、ユーザーの方がきっといろいろ出てこられるでしょうから、結果的に広く利用されることをもちろん希望しておりますが、前回は本当に、日本語に関するコーパス言語学というもの自体がジャンルとして存在しなかったということなので、いろいろ頑張ってそれを作ろうとしたわけですが、今回は幸いなことに定着しておりますので、そっちの方では余り、もちろん高度化していかなければいけないんですけど、だけど、立ち上げということは今回は不要かなと私は考えております。ひょっとしたら、ほかの人間はまた考え方が違うかもしれません。

○石川副主査

ありがとうございます。資料にあるロードマップ案で、3年目から予定されている「イベント」という辺りが、もしかすると、研究的なプラットフォームになる可能性もございますね。

○前川委員

それはそのような御理解で問題ないと思います。我々の方から進捗状況を報告すると同時に、新しく利用できるようになっておりますので、それを利用していろいろと、BCCWJ1との比較というのが最初行われると思いますけれども、スマホがいつから出てくるかとか、そういう話をいろいろとしていただければと思っておりますが、これについても、もし詳しい計画、山崎、小木曽の方で作っておりましたら、補足してもらおうと思います。

○石川副主査

どうもありがとうございました。大変、これからの日本語研究をまた大きく変える可能性のあるBCCWJ2、非常に期待が持てるところであります。御報告ありがとうございました。

○前川委員

ありがとうございました。

○相澤主査

ありがとうございます。

ほかはいかがでしょうか。　（挙手なし。）

では、また後ほどまとめて議論の時間を設けることといたしまして、次の発表に移りたいと思います。続きまして、私の方から「大規模言語モデルと言語資源」についてお話をさせていただきます。資料はこちらで共有したいと思いますので、しばしお待ちください。

では、始めます。よろしくお願いいたします。本日の話ですけれども、大規模言語モデルは非常に変化が速く、日々技術も変わっていきます。その中で、本日は比較的新しい論文ベースでトピックを御紹介をして、それを通して、大規模言語モデルでもよく使われる「コーパス」という言葉が、言語資源における「コーパス」と違うものなのか、あるいはどこが共通なのかということを議論する、そういうきっかけになってくれればと考えています。大きく本日の話題は二つございまして、一つは事前学習と呼ばれる、モデルの訓練に使われる言語コーパスの話、2番目は言語モデルが生成するテキストの話です。

まず、1番目、大規模言語モデルの構築、訓練に使われるコーパスの話になります。

こちらは、よく私が参照するlifearchitect.aiというサイトで、日々アップデートされている、今の言語モデルの一覧表となっています。新しいもの順に並んでいまして、1か月たっただけでもこの一覧表の最新モデルががらりと変わるというのがまた変化の激しさを物語っています。この表の各行がそれぞれ異なるモデルに対応していて、その中で、Tokens trainedと出ているものが、これらのモデルを訓練するために使われたトークン、つまり、ほぼ単語、言葉のユニットの数を表しています。御覧いただければ分かりますとおり、一昔前、例えば1年前と比べると、訓練に使われるトークン数、テキストの量というのは更に増加してきています。同じサイズのモデルを訓練する場合でも、より多くのテキストが使われるということになります。このTokens trainedというところに書いてある（B）というのは、Billion、つまり、これが10億の単位なので、18000と書いてあるということは、10億×1万8,000語という形になります。つまり18兆で、非常に多くのサイズのテキストが訓練に使われているということをまず御紹介させていただきます。

この中で、最近出てきたFineWebコーパスと呼ばれるコーパスがありまして、そのコー

パスがどのように構築されたかということを、かなり細くなりますが、本日は具体的に御紹介したいと思っています。このFineWebコーパスというのはどういうコーパスかというと、ウェブをベースにしたコーパスで、誰でもダウンロードして使えるようになっています。ここにあるとおり、“15 trillion tokens of the finest data the web has to offer”なので、非常に良質のウェブコーパスであるということがここで重要なポイントとなっています。FineWebがファインであるために、どう良質にしていったかという話がここでのトピックとなります。

では、この15Trillionトークンというのはどれぐらいかという話ですが、Trillionなのでゼロが12個、つまり15兆トークンです。どうも調べてみると、大体2億冊の本に相当、国会図書館だと12館分。恐らく、これまで出版された本全てよりも多い量のテキストとなるかと思います。先ほどのBCCWJ1は1億語（短単位）ということで、非常にアバウトですけれども、これが大体1.5億文字－トークンだったとすると、大体10万倍のサイズということになって、それからしても非常に大きな量のテキストが使われていることになります。

このテキストそのものはどこから出てくるのかというと、ウェブ上に公開される文書を寄せ集めてきたものから、それを洗練して作っていくということになります。文書を収集しているのは、FineWebではなく、Common CrawlというNPOです。Common Crawlは研究及び分析の目的でインターネット（ウェブ）上のコピーを無償で提供する非営利団体となっていまして、ウェブページ上で、このように年別に統計データを出しています。その累計はページ数で11乗。1ページ当たり、テキストの長さはまちまちですが、ある一定長のテキストがありますので、それなりの分量があるということになります。毎年分を累積していくと、当然のことながら、テキストの分量はどんどん増えていきます。逆に、かなり網羅的にウェブテキストをクロールしていますので、このFineWebあるいはCommon Crawlベースで使っているデータ以上に増える余地は余りないと。前回発表したとき、ウェブ上のテキストをほぼ使い尽くしているという表現をしたと思いますが、その根拠というのは、このように、網羅的にクロールされたリソースを全て使っていることを意味しています。

ここで問題となってくるのが、ウェブ文書を我々が見るときはテキストのように読んでいるかもしれないんですけれども、機械が読むウェブ文書というのはテキストではないという点です。左側の図は、非常に小さいですけれども、見た雰囲気で察知していただければと思います。ウェブページの中には、ブラウザ上ではこう表示するんですよといった指示タグが一杯付いていて、読むテキストというのはそのうちのごく一部になっています。なので、ウェブページから人間が認識する自然言語の文を抽出するためには、様々な処理をつないだパイプラ

インが必要で、それがすなわちウェブ文書のクリーニング、テキストの質を高めていく処理の流れというわけです。

FineWebの構築では、したがって、何をするかというと、ウェブの文書そのものから始まって、いかにそれを人間が使うテキストに近づけていくかという処理をすることになります。その背後にあるモチベーションは、大規模言語モデルを構築するに当たっての経験則あるいは直感というもので、そこに書いてあるとおり、大規模言語モデルの性能は、その事前学習に使ったテキストの品質とサイズに大きく依存するということがあります。サイズに依存するというのはこれまでもよく知られていたことですけれども、サイズだけではなく、品質が良い方が言語モデルの性能も良くなるということの重要性が認識され、そのために、ウェブからテキストを抽出する場合も、なるべく品質のいいものを作ろうということになります。

じゃあ、この品質というのは何か、品質というのはどうやって測るのかというのが次のスライドになります。コーパスの品質の測り方、言語学的にはいろいろなものがあると思いますけれども、大規模言語モデルを作る人はエンジニアやコンピューターサイエンスの人でして、言語そのもののクオリティーを人手で測るということはもはや余りしていません。ここに書いてあるとおり、問題を解かせてどれぐらい言語モデルが正解できるかを調べて、モデルではなくコーパスの品質を数量的に測るということをします。ここで、例えばChatGPTクラスの大きなモデルをゼロから学習すると、それだけでものすごくコストが掛かるので、もっとシンプルな小さなモデルを使って、同じ構成のモデルについて、サンプリングしたコーパス1で訓練したもの、別のコーパス2を与えて訓練したもの、二つを構成します。そうすると、訓練で使ったコーパスだけが違うモデルが2個できることになります。そこで、そのモデル二つに同じタスクを解かせて、その正解率を調べるということをします。このタスクと言っているのは何かというと、スライドの右側に一つだけ例を示しているとおり、例えば“What is the second most common element in the solar system?”という質問をしたときに、答えを“Iron”、“Hydrogen”、“Methane”、“Helium”から選びなさいと、この4択問題、この場合は知識問題ですけれども、こういった問題をたくさん解かせて、正解率を調べて、正解率が良い方がコーパスが良いと判断するというのが、ここでのテキストのコーパスの品質の測り方となります。

そういった準備を整えた上で、さて、では、品質の良いコーパスをどうやって構築していくかという処理の流れを示しているのがここのスライドになります。幾つか知られたクリーニングあるいは品質向上の方法というのがありまして、それぞれについて幾つか知られた方法

あるいは提供されているライブラリーというものがあって、それらが選択肢になるので、一体どれがいいのか、あるいはどのようにすればいいのかというのをステップ・バイ・ステップで地道に積み上げていくことをします。最初に適用するのは、ウェブ文書からテキストを抽出する処理、すなわち、先ほど見たようなたくさんコードを含んだウェブ文書から文だけあるいは文字列だけを抽出するという処理です。2番目は、URLベースで怪しいものを取り除く、あるいは、ターゲットとする言語が英語であれば、英語だけを抽出する、あるいは、過度に長いものはちょっと怪しいので取り除く、非常に短いものは文としての意味がないので取り除く、無意味に繰り返しているものは取り除く、そういったフィルタリング処理となります。3番目は重複除去で、テキストの重複が大きいと性能が落ちるということが知られているため、同じような文書を削除する処理です。重複というのは、この場合、全く同じ文書ではなく、ほぼ同じ文書を指していて、曖昧な判断が必要となりますので、かなり計算的にも重いですし、何を重複とみなすかという基準のところでパラメーターの調整が必要になるという、ノウハウが重要なものになっています。最後に、それでもちょっとおかしな文章は残るので、経験則に基づくフィルターを適用するという処理が入ります。こういった処理を積み重ねて文書をクリーニングしていきます。

それがいかに地道な努力の積み重ねであるかを示しているのが、FineWebの論文から引用してきた、こちらのグラフになります。例えば第一の処理であるテキスト抽出について言えば、二つ比べてどちらがいいかを比較したのがこのグラフで、横軸がTraining tokensなのでトークンの数、縦軸が先ほど御説明した問題セットの正解率となっています。そこで、二つのコーパスを比べて、青い線の方が品質が良くなっていると判断するわけです。その後に2番目の基本フィルタリング処理に行くと、今度は、赤い線の方がいいから、じゃあ赤にしようと。重複除去は様々な手法があるんですけれども、じゃあ、こちらにしようというように決めていきます。最後に幾つか追加フィルタリングという処理を適用して、最終的に、この場合はFineWebというのはオレンジなので、オレンジがベストな組合せになったと、そういうことをしています。

FineWebをベースに、更なる品質向上を目指したコーパスとしてFineWeb－Eduというものも作られています。これはFine Webの15兆トークンのコーパスから教育的価値の高い文書だけを集めて1.3兆トークンのサブコーパスを構築するという処理です。右側にあるグラフを見ていただければ分かるとおり、教育的価値の高い文書を集めたサブコーパスで訓練したモデルでは非常にタスクの性能が上がっています。つまり、知識問題、例えば先ほ

どのどの元素が多いですかと、そういう問題に関して言えば、教育的価値の高いコンテンツに重きを置くことでタスク性能は上がるということです。こういった形でタスクに合わせてサブコーパスを作っていくということも有効であることが分かります。

ここで次に御紹介したいのは、そのための手順です。普通の言語学の感覚で言えば、教育的価値の高い文書を選ぶというのは、すなわち、アノテーションする人間が教育的価値が高いという判断を下すことであると思いますが、ここでのやり方ではそうではなく、言語モデルに人間の代わりをやらせるということが行われます。つまり、対象とする文書は非常に多い一方で、言語モデルの性能が上がっていますので、アノテーションのような作業も、人間ではなく、言語モデルが当たり前のように開発の現場で使われているということです。具体的には、まず、高性能な言語モデルを持ってきて、その言語モデルにサンプルしたウェブページの教育的価値をゼロから5のスケールで自動評価させる。その自動評価させた結果を使って、もうちょっと安価に計算ができる、つまり、大量の文書を評価しなければならないので、計算コストが高い高性能モデルではなく、軽量な分類モデルが必要で、その軽量分類モデルの学習に、高性能モデルがアノテーションした結果を使うという形です。そうして、コーパスの全文書に教育的価値の有無についてのラベルを付けるという手順を適用しています。

更に話が細かくなりますが、教育的価値を判定するために、LLMに与えるプロンプトだけは人間が書いていて、そのプロンプトの例がここにあります。「ウェブページからの抜粋です。5段階評価システムを使用して、このページが小学校から中学校までの教育現場で教育的な価値が高く、教育に役立つかどうかを評価してください。各基準を満たすごとにポイントが加算されます。」ということで、様々な基準が細かく書いてあります。通常なら人間のアノテータに与えるべき指示ですが、この場合は言語モデルに対して指示としてこれを与えて、抜粋を検討した後に、「根拠を100字以内で簡潔に説明してください。最後にスコアを示してください。」と、そういうことまで自動化してやらせます。LLMを使ったこういうスタイルが当たり前になっているということで紹介をいたしました。

ついでながら、先のスライドで軽く補足してあったのを飛ばしてしまったのですが、追加フィルタリングのルールというのは非常にアバウトで、私も今回論文をじっくり眺めて驚いたんですけれども、文末にピリオドがないものは除くという、アバウトな処理がFineWebの一世代前のコーパスでは行われていたようです。FineWebではそれをもうちょっと賢くしたということなんですけれども、非常に進んでいる一面で、非常に大ざっぱなルールも適用されているということが分かります。

さて、細かい話はこの辺りにして、続きまして、言語モデルの学習の流れを見ておきたいと思います。言語モデルの学習というのは、大きく3段階で行われるとみなされています。3か月後にはもしかしたら変わっているかもしれないんですけれども、とりあえずは、まず事前学習、すなわち巨大コーパスを使うような学習、その後に中間的な学習があって、さらに最後の仕上げ学習があるという、この3段階です。この事前学習では、先ほど御紹介したFineWebなどが使われますが、多くの場合、単一のウェブコーパスをそのまま使っておしまいでなく、複数のコーパスを集めて、それらを学習したいモデルのイメージに合わせて混合して使っています。

このスライドは国立情報学研究所で構築している事前学習用のコーパスの変遷を示しています。サイズも増えていますが、あわせてコーパスの成分もかなり変わっています。最初は日本語をかなり重視していたのが、一番最新のコーパス、19.5兆トークンのコーパスですと、もう日本語は足りないこともあり、英語がかなりの割合を占めるコーパスを作成して学習に使おうとしています。その内訳を示したのがこちらのスライドになっていまして、先ほどから出ているFineWebがトップに現れています。正確にはFineWeb－2で、FineWeb自体は英語バージョンですが、FineWeb－2は多言語バージョンです。そのほかにも、様々なコーパスを組み合わせたりしているということが言えると思います。

そうした事前学習の後に、今度は単なるテキストではなく、合成したテキストで少し構造を持ったものを使ってモデルを賢くしていく。あるいは、少しドメインに特化したテキストを使う、さらに、ここにあるように、QAの形に成形した形のものを使うなど、目的に合わせてカスタマイズしたテキストで更に訓練する中間学習というものが入ってきます。

最後に事後学習として、人間が「望ましい出力」を指示することでモデルをチューニングする過程が入ってきます。これには、例えば要約してくださいと言ったら要約を出力する、あるいは、多肢選択問題を出せば選択肢を答えるといった指示への追従を教え込むものもありますし、右側にあるように、アラインメントと呼ばれる、社会的基準、倫理的基準あるいは法律を犯さない、そういったスタンダードに合わせるためのチューニングというのもあります。そうやって3段階の調整を重ねた仕上げとして、最後に大規模言語モデルができていくことになります。

再度確認ですけれども、事前学習はウェブから持ってきたものをクリーニングしていく。そのクリーニングの過程では、コーパスの品質を上げるためにかなりの選別が行われる。例えば有害コンテンツも除きますし、日本語らしいものを取り出すというフィルタリングも掛か

るので、フィルタリングのために参照した日本語、多くの場合、ウィキペディアだったりするんですけれども、結果として、目的に合ったテキストを恣意的に取り出すことになります。その意味で、コーパスは決してバランストではなく、質の良いものを意図的に選別していくという操作が入っています。ミッドトレーニング、中間のところでは、今度は、大規模言語モデルが出力したテキストを使うということにすると、そのモデルの出力に対するバイアスが掛かる。ポストトレーニングのところでは、人間にとって望ましいものを出させるようにするということで、社会的な基準にアラインメントするというバイアスあるいは人間の指示に従うというバイアスが掛かってくる。決して訓練に使われるコーパスはランダムなコーパスではないということには注意が必要になります。

そうやって訓練した大規模言語モデルはどのような出力を出すかという分析に関する御紹介が後半の部分になります。一体どんなテキストを出力するのかと。言語的に人間の出力と近いものが出力されているのかということになります。

ここで、大規模言語モデルの「口癖」というか、特徴に関する分析結果が最近の論文で報告されていますので、御紹介します。この論文でやっているのは、生成テキストから大規模言語モデルのどれが使われたかを特定できるかということです。例えばChatGPT、Claude、Grok、Gemini、DeepSeekの五つを持ってきたときに、あるテキストがこの五つのうちのどれからの出力かを言い当てられるかという、そういう話であります。

それを絵にしたのがここにあります。同じプロンプトを与えておいて、それぞれ五つのモデルに答えさせます。モデルを言い当てるに当たっては単純な方法を使うのではなくて、ちゃんと答えられるように訓練をした自動分類器というのを作成して、その分類器がどれくらい正解できるかを調べます。

その結果がこちらの識別性能とあるものです。512トークンを読んだ後の識別性能は実に97%。つまり、この五つのモデルは、512トークンぐらいを見ると、97%の正解率で当てられる。ランダムに選べば20%なので、97%―非常に高い正解率で答えられてしまうというのがその結果になります。最初の1トークンだけでも50%の正解率なので、これらの言語モデルは相当な個性を持っているというようにも言い換えることができるかと思います。

では、その個性はどこから来るのか、一体どうしてこんなに分かるのかということで幾つか実験がなされていて、まず、語順をランダムに入れ替える、つまり、それは言語でも何でもなくなってしまって、語の分布しかなくなるわけですが、語順をランダムに入れ替えても実はかなり分かってしまうというのがこのshuffling　wordsの結果で、chat、instruct、bas

eというのはそれぞれモデルを表していて、chatが先ほどの5分類だと思っていただくと、chatの場合、90%ぐらい出てしまっている。つまり、語の分布そのものが違うということが分かります。実際に特徴的なフレーズというのがそれぞれあって、例えばCertainlyなど。英語なので日本語にそのままマッピングするのは正確ではないですが、例えば「確かに」とか「承りました」とか、いろいろ口癖を持っている。先ほどのモデル構築過程で触れたチューニングの際に、このように言いなさいと教え込まれるために出てきてしまうような口癖もありますし、あるいは、言語モデルでよく見られる箇条書での出力について、箇条書の番号が「1.」であるか、あるいは「(1)」であるか、それぞれ傾向が違うので、そこら辺で特徴が出てくるというのは納得できると思います。

しかしながら、意味を変えないように言い換えしてくださいということで調べてみても、性能は余り落ちない。つまり、出力内容についても個性を持っているということもまた言えそうだとの指摘もされています。試しに－人間ではなくて－LLMに説明させてみると、こうしたことをLLMに説明させるというのは、言語学の人から見たら恐らく許せないかもしれないところですが、こうなります。つまり出力をChatGPTに与えて、一体この出力はどういう特徴を持っていますかを尋ねると、例えばChatGPTなら、Descriptive and Detailed Toneであって、Specific and Technical Wordが入っている、Claudeであれば、Concise and Straightforward Toneであるとか、人間がそう思うかはともかく、少なくともChatGPTの判断によると、それぞれの言い回しにも特徴があるとの判断です。なので、単に表層的な違いだけではなく、スタイルにも違いが生じていると分析される、それがこの結果でありました。

LLMの構築・推論過程のどこでこういった出力の特徴が付与されるのかを調べるため、オープンなモデルを使って、異なるモデルを同じ合成データで訓練して比較すると、どうも出力特徴は似てくるという結果も得られています。これによって、中間とか事後の訓練のところで特徴が出てくるのだろうという推察はできますけれども、恐らくそれだけではないだろうとも指摘しています。逆に、出力テキストの近さからモデル同士の類似度が分かるかというと、それはかなり分かるということで、出力からモデルの来歴、すなわちどのようなデータで訓練されたかはある程度推測できそうだという示唆も与えられています。ただし、これは、LLMが人間に似た文章を生成できないと言っているわけではなくて、人間に似た文章を生成してくださいというように指示をすればもちろんまねはできるんですけれども、通常の対話形式そのままの出力の形だと人間に非常に似たものを出すということはしていないと解釈できると

思います。する理由がないというのが、私の推測です。一つは、人間は危ないことも言いますし、文法的にも間違えたものを出しますし、そういったことまでまねする必要はないし、社会的に見て言語モデルとしては許容されないということもあります。もう一つは、言語モデルの出力を競合他社が開発に使うことを禁止している商用モデルもあり、特徴を持たせておくというのは予防的な意味で、実際プロバイダーにとっても役に立つ側面があると言えるという点もあろうかと思います。これは飽くまで推測ですけれども。

逆に、じゃあ人間は分かるのかということで、LLMによって自動生成されたテキストを人間は識別できるのかについて、これは先週開催されたACLという自然言語処理の国際会議での発表からの報告を紹介します。これはLLMの出力を見たことがない人は実は見分けることができないんですけれども、日頃接している人はかなりの高い精度で、しかも、ここにあるように文単位でLLMと人間の出力が見分けられるというのが結果であります。

他の論文で紹介されている結果も、現在のところは同様の報告をしていまして、例えば人間やLLM同士の識別の度合いを、クラスタリング（clustering）して視覚化したこの論文でも違いが明確に出てきます。これら全てに共通することとして、どうもLLMと人間の生成するテキストというのは、かなり個性を持っていて、区別はできるらしいということであります。

最後に、AIが生成するテキストの論文への影響の分析にいついてです。じゃあ、そうしたAIが生成するテキストというのは人間の言語にどういう影響を与えるかということで、これは前回も少し御紹介したかと思うんですけれども、特に論文を書くときには生成AIをよく使う。つまり、英語で書いたものを生成AIにプルーフリード（proofread）してもらったり書き換えてもらったりすることは最近よく行われるようになっています。それが一体論文の英語にどのくらいの影響を与えたかと、そういう調査というのは幾つかあって、その例がこのグラフとなっています。特に有名なのはDelveで、Delveという単語が、ChatGPT登場以来、急激に増えたというのは有名な話であります。それに限らず、生成AIテキストを特徴付けると言われている語はたくさんあるんですけれども、そうした語を手掛かりにAI生成テキストを自動分類しようとすると必ず、人間が書いたのにAIが書いたと判定されるものが出てくるという結果も報告されています。つまり、この場合には自動判定では、ぬれぎぬを着せられる人というのはどうしても出てくるということであります。

こうしたDelve問題について詳細に分析した論文というのがありましたので、こちらのスライドで紹介しています。先ほど見たとおり、“Science English”というのは、生成モデルの影響を受けて急速に変化しているものであると。その理由として、言語モデルが、「lexi

cal overrepresentation」と呼んでいる、普通よりもずっと高い頻度である単語を出してくる現象があり、それによって、特にノンネーティブ・イングリッシュ・スピーカーは影響を受けてしまう。その「lexical overrepresentation」の原因が一体モデルのどこから来るのかというのを調べようとした論文です。それで、いろいろやってみても外からは分からないというのがここでの結論であります。特にDelveは、アノテーターの偏りによるものだという説が非常に強力であったのですが、中身が分からない状態で、それを科学的に実証することはできなかったというのがポイントです。例えばナイジェリア・イングリッシュのような、そういうイングリッシュ別の語の頻度を持つようなコーパスを使って比べてみても、特にナイジェリア・イングリッシュにDelveが多いという証拠はないということで、Delveがどこから来たかについては、モデルを作った人が透明性を持って、こういうデータで訓練しましたということを開示しない限りは分析できないと。つまり、言語の解析のためにモデルの透明性が求められていくというのが結論であります。

今日のまとめについて、ちょっと散漫になってしまい、どうまとめるか難しいところですが、言語モデルが生成する言語というのがあふれ、それが人間の言語にも影響を与えている現状がある一方で、言語モデルの生み出す言語を本当に解析しようと思うと、言語モデル自体の透明性というのが求められてくる。透明なだけではなくて、実際にどのような言語を生成しているかをモニタリングすることも必要になってきますが、実際問題として、そこまでオープンなモデルというのは本当に数えるほどしか存在しない、モニタリングも難しい、というのが今後の課題になってくると考えています。

以上で本日の発表を終了いたします。どうもありがとうございました。

司会に戻りまして、質問がありましたら、どうぞよろしくお願いいたします。この後、討論の時間がございますので、討論と併せてしてもらってもいいですが、もし発表についての御質問がありましたら、よろしくお願いいたします。

○石川副主査

よろしいでしょうか。石川です。

大変に啓発的なお話、誠にありがとうございました。生成AIをめぐっては、本当に短い期間で、状況が大きく変動しているということを強く感じさせられる御報告であったかと思います。幾つか事実確認的な質問でございます。

1点目ですが、Common Crawlという団体に関して、「ネットの文章のコピーを無償で提供する」とございますが、これはどういう権利関係になっているのでしょうか。自身が作成したものでないテキストを第三者に提供する、という辺りに素朴な疑問を持ちました。

○相澤主査

そうですね。これは先ほども話題に出ましたが、いろいろな解釈があるところだと思います。Common Crawlに関しては、現実問題としてはリソースとして広く受け入れられているものになっています。日本に関して言えば、ウェブから機械的に収集してきたテキストを機械処理の目的で使うということについては、問題はないとされていますが、ただ、中身が全部分かるわけではありませんので、それについては使う人が注意していくことになると思います。

○石川副主査

ありがとうございます。

2点目ですけれども、データの収集に関して、ウェブ基盤の大規模コーパス構築などでは、内容的なバランスへの配慮としては、例えば、名詞リストなどからランダムに選んだ語をシーズ（seeds）として検索サイトに送り、データを収集するというようなことがなされていたと記憶していますが、同様のプロセスはあるのでしょうか。

○相澤主査

いえ、バランスを取るという操作は入っていないですね。

○石川副主査

なるほど。

○相澤主査

あるものは全て使うというのが原則で、もしかしたらバランスを取ったら質が良くなるかもしれないんですけれども、私の知る限りでは、クリーニングはするけれども、そういった意味でのバランスは……。

○石川副主査

バランスを取るという意味でのフィルタリングは通例掛けないということなのですね。

○相澤主査

はい。

○石川副主査

分かりました。ありがとうございます。

3点目です。今回御紹介いただいたFineWebの、“fine”というのは、クリーニングプロセスが入って重複が除かれているため、ということでしょうか。

○相澤主査

質が良いということだと思います。実際に彼らが示している最後のこのフィルタリングをした後の性能が基礎のものよりも上であるということです。一番右のオレンジのところが“fine”の由来だと思います。

○石川副主査

ありがとうございます。なるほど。FineWeb自体もよくできているわけですが、教育的テキストに限ったFine Web-Eduになると更に良くなるというわけですね。

拝見した資料でも、Fine Web-Eduの性能はFineWebを圧倒しているようで、教育的なテキストの価値を示す結果かと思います。BCCWJ2でも教科書データの収集を進められると伺ったところですが、生成AIの世界で、教育的なテキストに限って集めていくというような動きは広まっているのでしょうか。あるいはこの例に限ったことなのでしょうか。

○相澤主査

質が求められるというのは、やはり質が良いものを増やしたいという強いモチベーションになっていますので、一般的な取組です。特に今注目されているのは数学問題で、数学関係の文書を集めたり、数学関係の問題・回答を合成したり、あの手この手で集めて、モデルの推論能力を高めようという試みがなされています。ただ、全体としてはやはり量が必要なので、全体を集めながら、その後の中間学習と言われていた部分で例えばより質の高いも

のを学習させるとか、本当に職人芸に近いところでバランスを取っていると思います。

○石川副主査

ありがとうございます。

最後、４点目の質問でございます。資料の４ページ目で、最近の言語モデルの状況をお示しいただいているのですが、色の塗り分けは国と関係しているのでしょうか。

○相澤主査

すみません。私、ちょっと色分けの意味までチェックしておりませんでしたが。

○石川副主査

ざっと拝見すると、中国などの企業名も目立っているようですが、こうした言語モデルの開発競争において、日本の位置付けというのはどのようなものなのでしょうか。また、この数年で位置付けに変化があったとすると、どのような変化となっておりますでしょうか。

○相澤主査

日本もかなり頑張って追従しようとしていると思います。日本も有名なモデルが幾つかあるんですけれども、日本語版の性能ではかなりこの辺りのモデルに迫るところまで来ている場合もあります。ただ、まだユーザーへの親切さみたいな、サービスをする際のこなれ方の問題は残っているというのが、私の個人的な感じです。

○石川副主査

日本が得意そうな部分が、実は苦手だということでしょうかね。

○相澤主査

経験だと思います。

○石川副主査

なるほど。

○相澤主査

テストの問題で、クエスチョン・アンサーに答えるというところではかなり遜色のないところまで来ていると思います。あとは、サービスの実経験が増えるとか、データをもっと蓄積していくとか、そういったことはやっていく必要があると感じています。

○石川副主査

ありがとうございました。大変勉強になりました。

○相澤主査

前川委員、よろしくお願いいたします。

○前川委員

私、最近いろいろこういう観点の知識をちょっと勉強したいなと思っておりましたので、大変勉強になりました。ありがとうございました。

それで、2点伺いたいんですが、そのうちの1点は、最初の方にコーパスの性能の評価の話が出てきて、やはりかなりヒューリスティック（heuristic）な形での評価が行われていると－最後に名人芸というようなことをおっしゃいましたけれども、それに近いのかなという気がするんですが、その中で、今回例として挙げておられたのが、今出ていますね。太陽系で2番目に多い元素は何かというような、こういう質問に関する正解を与えられるかというような観点で評価するという例だったんですけれども、このように客観的な知識ばかりなんですか、それとも、もうちょっと何か……。

○相澤主査

非常に評価用セットは多様なので、例えば、ある文とある文が、ある文から別の文が推測されるかというナチュラル・ランゲージ・インファレンスみたいなものも含まれていますし、あるいは、要約するとか翻訳するといったタスクも考えられます。ただ最近のチャレンジは知識問題になっていて、例えば大学院の博士課程レベルの問題を集めましたとか、そういうかなり難しい知識・推論系が多いのは事実です。ここで、知識と言語的な能力を分けて評価できればいいんですけれども、両者は割と分けるのが難しいので、どうしても一緒に評価することになり、こういうタイプのクエスチョン・アンド・アンサリングなど機械的に評価ができる設定

が多いと思います。また、対話的な能力の評価は別に行われて、二つのモデルからも出力を提示して1対1形式でどちらがいいですかというのを、LLMが判定するという方法でやられたりします。

○前川委員

分かりました。評価である以上、ある程度客観的な評価ができないとしようがないですから、どうしても論理的なものに偏るということなんでしょうね。

それからもう一点、コーパスの構築というか、その中で重複排除というのがありましたね。あれ、実は我々、昔、日本語ウェブコーパスというのを作ったときにやはりやりまして、そのとき一番多い文は何だったかというと、「文書の先頭へ戻る」というのが一番多かったんです。圧倒的に多くて、それは一文として、重複は数えないんですけれども、ここで排除している対象というのは同じ文書なんでしょうか。それとも、文あるいは句、フレーズみたいな短いものなんでしょうか。

○相澤主査

すみません。私、文書と書いてしまったんですけれども、文章ですね、いわゆるセグメント。

○前川委員

もっと短い、文とかフレーズとかですよね。

○相澤主査

短い単位です。はい。

○前川委員

分かりました。以上です。

○相澤主査

ありがとうございます。

では、質問も踏まえながら、そろそろ意見交換に入っていきたいと思いますので、よろし

くお願いします。

本日は、現代日本語書き言葉均衡コーパスについてと大規模言語モデルについてということで御紹介させていただきました。本日の二つの発表は、お手元の参考資料2の中の前期の報告のフォローアップに当たるものということになります。本日の発表に限らず、委員の皆様の御専門分野と絡めながら、もしお感じになられることや御意見、御感想などありましたら伺ってまいりたいと思います。また、山崎先生や小木曽先生も是非御発言を頂ければと思いますので、後ほどお声掛けをさせていただきます。

まず、追加で何か御意見、御感想等ありましたら、よろしくお願いいたします。

○石川副主査

よろしいでしょうか。本日は国語研から山崎先生、小木曽先生がおいでくださっております。実際の作業を担当しておられる両先生から御覧になって、BCCWJ2の特性であるとか新しさであるとか、あるいはBCCWJ1と比べた場合の変化など、コメントを頂ければと存じますが、いかがでしょうか。

○相澤主査

では、山崎先生、小木曽先生、それぞれから一言ずつ頂ければと思いますが、よろしくお願いいたします。

○山崎客員教授

では、僭越(せん)ながら。山崎です。私からちょっと。

今、正に2011年とか12年とか、そういうデータが大量に届いているところで、それを所内でいろいろなチェックをして、データがちゃんとなっていればXMLのデータベース格納に行くんですけれども、なかなかデータそのもののチェックというか、どうしても人手で見なければ分からないエラー－これはOCRのエラーが多いんですけれども－その辺のところの確認で非常に時間を取っているというのが率直な感想です。データの作成は、外部の業者に頼んでいるんですけれども、なかなか99.95％という精度を達成できるところは難しいというのが悩みの種であります。

○石川副主査

ありがとうございます。

○相澤主査

ありがとうございます。

では、小木曽先生、よろしくお願いいたします。

○小木曽教授

そうですね。基本的にBCCWJ1の後の時代のものを追いかけている形で、バランスの取れた－書籍には主に限られるんですが－そういったものが出来上がりつつあると思います。形態素解析の精度などを主に昔から担当してきていたんですけれども、以前、BCCWJ1のときはゼロから作らなければいけなかったのに比べて、今回はほぼほぼできたものがあるのをブラッシュアップしていけばいいということで、精度は相当に良いものになるので、日本語研究にとってはかなり使えるものになるのではないかなと思っています。あいにく、先ほど前川の方から話があったように、体制が大分変わってしまっていまして、所内で直接対応できる人間が、専任では、私一人みたいな状況になっていますので、あとは特任とか非常勤でお願いしている感じなので、なかなか以前のような形にはできないというところはあるんですけれども、何とかいいものにしていきたいなと思っています。先ほど石川副主査から、イベントなどで以前のような学術的盛り上がりという話もあったんですが、この辺もちょっと今のような難しさはありまして、委託研究ということで、データを作る資金を頂いているわけですが、研究にと盛り上げていくというところはなかなか、そっちではできないものですから、ほかと組み合わせてできることはやっていきたいなと思っているところなんですけれども、日本語研究にとって使えるものというのは最低限のものとして作れるように頑張っていきたいと思っております。

○石川副主査

ありがとうございました。

○相澤主査

ありがとうございます。

ほかはいかがでしょうか。　　（挙手なし。）

では、そうしましたら、まず会場にいらっしゃる神永委員から順番に、武田委員、森委員の形でお伺いさせていただければと思います。

○神永委員

神永です。今日は本当にいろいろと貴重なお話をありがとうございました。

先ほどもBCCWJ2のところのお話で申し上げましたけれども、非常にこれ、BCCWJ2コーパスは国語辞典の編集に間違いなく活用できるものでありますので、むしろ、活用するだけじゃなくて、これによって国語辞典の内容が大幅に変わるものだと考えていますので、2028年の完成、是非期待したいと思いますし、また、そのとき、ちらっとおっしゃっていました、公開方式、「中納言」を無償でオンライン公開を考えていらっしゃるということ、これも非常に有り難いことですので、是非お願いしたいなと思っております。

あと、相澤主査のお話なんですけれども、これも実は、私も辞書編集者の立場から考えてみると、非常に面白いと言いますか、興味深いことがございまして、すごく細かなことなんですけれども、何か口癖のようなものがあるとおっしゃっていましたけれども、これも非常に、今英語で実例をお出しくださっていますけれども、きっと日本語でも何かあるんじゃないかなと。それは何か辞書に反映させていくことになるかなというような気もします。

それから、今までは私たち、コーパスで収集した情報－用例なんですけれども、これを基にして、機能的と言うんでしょうかね、意味を記述していたんですが、逆に、今度、このLLMはそれをやってくれるわけなんですね、取りあえず。これは一体どのようなことになるんだろうかということもちょっと考えたいですし、それで出てきたもの－いや、実際にそれをやっているところも何かあるような気もするんですけれども－それが出てきたときに我々辞書編集者としてどうやって今後対応していったらいいのかということがあるかと思いますので、その辺のことも考えながらいきたいなと思っております。

今日はどうもありがとうございました。

○相澤主査

ありがとうございます。

では、武田委員、よろしくお願いいたします。

○武田委員

武田でございます。本日は貴重なお話、大変ありがとうございました。

BCCWJ2につきまして、若干質問になってしまうかもしれませんが、対象となるレジスターがBCCWJ1に比べて、新聞、雑誌等が対象とならないということで、大変残念に思うところがありましたが、それにつきましても、小木曽先生の昭和・平成書き言葉コーパスなどで部分的には検索可能ということで、若干安堵しております。

同様に、対象となるレジスターが大きい方がよろしいのかと思われますが、教科書については、科目は小・中・高全科目が対象ということですが、その中でも占有率の高いもの1冊を対象とするという形であるようですが、教科書なども非常に点数出ておりますし、それぞれ取り上げている教材も、例えば国語などではかなり幅があるかと思います。そういった形で、教科書の対象を1冊から広げるというようなお考えはございますでしょうか。

○前川委員

これは小木曽さんか誰か。

○小木曽教授

私からお話ししましょう。

教科書についてまずお話しするとしますと、こちらは、一つはBCCWJ1の時の設計を引き継いでやっていくということで、同じやり方で作るというのが今回の方針ではあります。ただ、BCCWJ1の時はサンプリングで、実は先ほどのスライドに出てきたもののごく一部しか出ていなかったものが全部出るようになるという点では多くなります。ただ、量の問題と言いますか、一番よく使われているもの以外に広げられるかという点ですけれども、この辺りはかなり予算と人手の兼ね合いがもう、ほぼそれで決まってくるような状況でして、御存じのとおり、昨今、あらゆるものの値段が上がっていますので、2億語にするというところは死守しなければならないと思って、それは出版サブコーパスの書籍の拡張で必ずやるということにしていて、残りの部分はできれば広げていきたい。今日お話があった教科書とSNSについては少なくともやることはできるというところをお話ししています。場合によって、余り可能性はないかもしれませんが、非常に効率よく進んだことで、ほかのことに手を出す余裕があるということになりましたら、BCCWJ1の時の非母集団という形でいろいろな種類のものを入れていたわけですが、あの部分を拡張するということもあり得ると思いますし、また、教科書についても、例えば国語の教科書については上位何社分まで広げるとかいうような可能性はな

いではないけれども、ちょっと今の状況を見ていると、簡単ではないかなと思っているところです。

○武田委員

分かりました。ありがとうございます。

○相澤主査

ありがとうございます。

では、森委員、よろしくお願いします。

○森委員

今日はありがとうございました。BCCWJ2でSNSが加わることによってどんな情報が増えていくんだろうということをすごく楽しみに思っております。

マスコミに身を置く者としましては、新聞が外れてしまったのがとても残念で、恐らく新聞が有料化してしまっているというのは、自分たちが取材したものについて、それを勝手に利用されることを防ぐためということなんですが、今回は情報の利用ではなくて、言語の方ですから、どうぞ是非使ってくださいと言ってもらえるような状況にならないものかと、テレビ局から何か言えるものではないんですが、同じマスコミとしてはそのように私は感じました。入っていないというのがとても残念です。

ただ、SNSはより今を映すものでしょうから、例えば2028年のものは2028年のものだし、2025年のものは2025年のもので、それを2030年に見たら、当時はこんな使い方をされていたんだなという感覚が、こちらのBCCWJ1のときに資料として使われていたもの以上に出てくるだろうなという気はしまして、それは恐らく更新していかなくてはいけないので、その環境が整うことも強く希望します。

それから、相澤主査からお話しいただきましたものもすごく面白くて、特に口癖に関しては、これ、日本語でもやれるものであれば遊んでみたいと、本当に純粋な興味、関心が湧きました。もし日本語で、例えば我々が書いたものであったり、しゃべったことであるものを分析されると、ふだん読んでいる新聞は何なのかとか、ふだん読んでいる文献、あとは書籍はどういうものなのかというものを恐らく見抜かれてしまうでしょうし、あとは、関西にルーツがあるんじゃないかとか、九州にルーツがあるんじゃないかということまでも分かっちゃうんだ

ろうなという、ちょっと面白さと怖さを同時に感じました。

ありがとうございました。

○相澤主査

ありがとうございます。

それでは、植木委員、よろしくお願いします。

○植木委員

ありがとうございます。本日は相澤主査、前川委員のお話、本当に知らない、全く門外漢として学ばせていただきました。ありがとうございました。

こちらはいろいろなコーパスも生成AIのようなものも利用するだけの立場なので、それを作り上げていくのにいろいろな御苦労や、精緻な段階を追ってされているんだなということを改めて教えていただきました。

相澤主査のお話で、本当に素人の感想のようなことなんですけれども、LLMによって作られたテキストが人間には識別できるのかというところで、案外できるものなんだという、生成AIに特徴的な癖がある、単語があったりする－今、口癖のお話も出てきたんですけれども、そこは非常に興味深く思いました。今ちょうど大学では春学期が終わってレポートなどが出てくるところなんですけれども、やはり学生たちが生成AIで作ったレポートを提出してくるということがもうこのところ出てきて、こちらも非常に気を付けて、それができないような細かい指示をした課題を出すんですけれども、その中でもやはり使ってくる。それが、経験的にやはり何となく分かるんです。例えば、私は古典なので、この世に存在しない和歌のような変なものが記されていて、何とかの何番の歌なんて書いてあるんですけれども、存在しない歌でありますし、今のところはこちらの方が分かっているので、「それはどこから取りましたか」ともう一度フィードバックすると、「すみません、生成AIで作りました」と返ってきます。今のところはそれができるんですけれども、すごくジレンマなんです。大規模言語モデルの学習用のテキストの品質を保持して、その精度が上がれば上がるほど、今度は例えば学生が生成AIで作ってきたものが非常に自然になり、内容も正しくなってきて、教育の現場でそれをどのようにしていったらいいのかということが大きな課題になってくるということを改めて思って……。この小委員会で対象にするところとちょっと違う、付随した別の問題ではあるんですけれども、教育現場における生成AIのようなものの取扱いというのも非常に難しい問題だと思

いながらお聞きしておりました。

ありがとうございました。

○相澤主査

ありがとうございました。

じゃあ、長い御発表を頂いた後ですが、前川委員もよろしくお願いいたします。

○前川委員

特に私自身の発表についてはないんですが、先ほど、森委員でしょうか。SNSの利用法みたいなことについてちょっとお尋ねに近いものがあったと思いますので、それについての考えを述べさせていただこうと思います。SNSも我々2種類のものを考えているんですけれども、そのうちのLINEの方に関しては、BCCWJ1と比べますと、明らかに違った性格を持っていると私は思っております。と言いますのは、LINEのチャットというのは皆さん、国民の91%が使っていますから、多分皆さんお使いだと思うんですけれども、あれは対話なんですよね。こっちが何か発信して、向こうから返ってきてと、そういうインタラクションが非常にはっきり表れるメディアですので、それは実はこれまで、話し言葉コーパスにはありましたけれども、書き言葉コーパスには、BCCWJ1にはその部分に該当するようなものはほとんどなかったわけです。その意味で新しい価値－サイズは小さいんですけれども－新しい使い道が特にそこには出てくるんじゃないかなと思いますし、ひょっとしたら大規模言語モデルのファインチューニングみたいなものにも利用していただけるのかなとも考えております。

以上です。

○相澤主査

ありがとうございます。

では、せっかくですので、山崎先生と小木曽先生も一言ずつ頂いてもよろしいでしょうか。山崎先生から。

○山崎客員教授

全体の感想ということでよろしいですか。

○相澤主査

はい。

○山崎客員教授

特に私、BCCWJ2はもう知っているので、あまり新しい情報はないんですけれども、相澤主査が御説明くださったLLMの現状というのは、新鮮な情報がたくさん多くて、このスピードで進んでいくと、最終的には人との区別がなくなるですとか、どこまで行くのかというのはちょっと末恐ろしいような印象を抱きました。ひょっとしたらどこかでストップを掛けるということもあり得るのかなという気がしています。

以上です。

○相澤主査

ありがとうございました。

では、小木曽先生。

○小木曽教授

私もやはり相澤主査のお話、興味深く伺ったところです。BCCWJの規模だと事前学習とかには全く及ばないというのは最初から分かっていたことではあるんですけれども、今後どういう使い道があるかというのを考えていくときに、一つ、今我々が使っている日本語が生成AIによって、ある意味汚染されると言うと悪い言い方過ぎるかもしれませんが、人間だけが作っていた言葉の時代の最後の日本語のデータ集合になる可能性があると思っています。我々が作っているのはほとんど紙の本から集めてきてやっていて、2025年までなので、ほとんどまだ生成AIが関与していない状態であり、そのようなちょっと違う意味でBCCWJというものの価値が見えてきて、ひょっとするとそれは今後、大規模言語モデルを評価したりするときに利用できるという新しい価値が出てくるのかもしれないなと思って拝聴しておりました。ありがとうございました。

○相澤主査

ありがとうございます。

石川副主査、どうぞ。

○石川副主査

ありがとうございました。たくさん質問をさせていただきまして、本当に勉強になりました。ありがとうございます。

BCCWJ2については、小木曽先生が最後におっしゃってくださいましたように、一般の言語使用に生成AIの影響が生じ始める直前期の貴重なデータとなるという点、これは盲点でした。新しいBCCWJ2には、そのような価値付けもあるということを改めて理解した次第です。

今日は、両先生、ありがとうございました。

○相澤主査

ありがとうございました。

それでは、ちょうどお時間も来ましたところで、本日、現代日本語書き言葉均衡コーパスの拡充が予定どおりに進んでいることが確認できまして、また、大規模言語モデル前の最後のコーパスとして順次公開されていくということは非常に楽しみであるということで皆様の意見が一致したと思います。生成AIで生成された文をいかに扱っていくかというのは、今後の急激な変化が続く中で引き続き見守っていく必要がある課題かと思います。特に皆様から追加の御意見ないようでしたら、最後に事務局から御連絡をよろしくお願いいたします。

○鈴木国語調査官

本日はありがとうございました。

事務局からの連絡でございます。

参考資料1の「当面の予定」、こちらにお示ししておりますが、次回の言語資源小委員会については、第9回、9月9日火曜日でございます。時間が本日とは違いまして、15時から17時の予定でございます。詳細は改めて御連絡いたします。

それから、もう一点です。今期の後半の言語資源小委員会ですとか国語分科会の日程を決めるための皆様の御予定を伺う用紙をお送りいたしますので、そちらの方、どうぞよろしくお願いいたします。

以上でございます。

○相澤主査

では、本日の言語資源小委員会はこれにて閉会といたします。どうもありがとうございました。