whole-NWJC: 『国語研日本語ウェブコーパス』全データ

浅原正幸　国立国語研究所・総合研究大学院大学　masayu-a@ninjal.ac.jp

概要

本稿では、『国語研日本語ウェブコーパス』（NINJAL Web Japanese Corpus: NWJC）の全データ（whole-NWJC）について概説する。本データは、国立国語研究所との共同研究を通じて利用できる。

1　はじめに

本研究では、国立国語研究所プロジェクト「日本語記述の緻密化を目指した超大規模コーパスの構築」において整備された『国語研日本語ウェブコーパス』(NINJAL Web Japanese Corpus: NWJC) 全体（以下、whole-NWJC）について概説する。『国語研日本語ウェブコーパス』は、2014 年第 4 四半期（2014-4Q）に検索システム「梵天」および「中納言」を通じて共有された[1)]ほか、語彙表および n-gram データの形式で公開されてきた。

本稿では、特に大規模言語モデルの構築に資するデータとして、2012 年第 4 四半期から 2015 年第 2 四半期に収集された WARC 形式のデータを新たに共有することを報告する。以下において、共有データの概要とその詳細について解説する。

2　『国語研日本語ウェブコーパス』既共有データ

表 1 に『国語研日本語ウェブコーパス』（NINJAL Web Japanese Corpus: NWJC）の既存共有データの概要を示す。本データの設計に関する詳細な情報については、文献 [8, 9, 10] に記載されている。

当初、本プロジェクトでは検索システム「梵天」[11, 12] を通じて、2014 年第 4 四半期（2014-4Q）のデータを公開していた。このデータは、収集 URL 数 83,992,556、文数（延べ数）3,885,889,575、文数（異なり数）1,463,142,939、国語研短単位数 25,836,947,421 を含む大規模なものであった。しかし、2021 年 12 月 24 日をもってサーバ維持費用の制約により共有が停止された。

さらに、検索システム「中納言」を通じて、2014-4Q のデータの一部（国語研短単位数 86,277,772、NWJC-2014-4Q の 0.33%に相当）を公開していたが、2024 年 2 月 29 日に共有を停止した。

これに加えて、統計情報として語彙表、中納言に搭載されたデータの語彙表、および n-gram データが公開された。

また、単語埋め込みモデルとして、NWJC2vec [1, 2] および chiVe [3, 4, 5] が提供された。さらに、事前学習済み BERT モデルとして、語彙素に基づき文単位で訓練された NWJC-BERT [6] と、文章単位で訓練された chiTra [7] も公開されている。

3　『国語研日本語ウェブコーパス』データ構築手法の概要

『国語研日本語ウェブコーパス』（NINJAL Web Japanese Corpus: NWJC）は、言語研究に資する言語資源として、ウェブを母集団とし構築された大規模なコーパスである。当初、本プロジェクトでは 100 億語規模のコーパスを構築し、外部研究者が検索システムを介して利用可能とすることを目標としていた。本節では、NWJC のデータ構築手法について概説する。

ウェブページの収集には、Heritrix クローラ[2)]を用いた遠隔採取（remote harvesting）を採用した。Heritrix は、Internet Archive によって開発されたウェブアーカイブ専用のクローラであり、各国の国立図書館などが国内外のウェブページを保存するために広く利用している。日本国内においては、国立国会図書館が実施するインターネット資料収集保存事業（WARP）[3)]においても用いられている。

Heritrix は、ISO 28500 で規定される WARC（Web ARChive）ファイル形式を使用してウェブアーカイブを保存する。この形式では、ウェブ資料を HTTP ヘッダやレスポンス情報とともに単一のファイルに格納する。データ処理には、Python の WARC ライブラリ[4)]を利用することで効率的な操作が可能である。

1）「梵天」は 2021 年 12 月 24 日に共有を停止、「中納言」は 2024 年 2 月 29 日に共有を停止している。

2）https://github.com/internetarchive/heritrix3

3）https://warp.da.ndl.go.jp/

4）https://github.com/internetarchive/warc

表 1　『国語研日本語ウェブコーパス』既共有データ

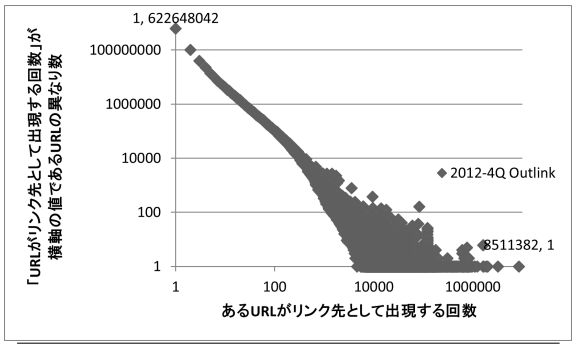
データ名	URL
「国語研日本語ウェブコーパス」(2014-4Q) 語彙表	https://github.com/masayu-a/NWJC
『国語研日本語ウェブコーパス』中納言搭載データ語彙表	https://doi.org/10.15084/00003666
「国語研日本語ウェブコーパス」n-gram データ	https://www.gsk.or.jp/catalog/gsk2020-c
「国語研日本語ウェブコーパス」NWJC2vec [1, 2]	https://www.gsk.or.jp/catalog/gsk2020-d
chiVe: Sudachi と NWJC による日本語単語ベクトル [3, 4, 5]	https://github.com/WorksApplications/chiVe
「国語研日本語ウェブコーパス」NWJC-BERT [6]	https://www.gsk.or.jp/catalog/gsk2020-e
chiTra: Sudachi と NWJC による Transformer モデル [7]	https://github.com/WorksApplications/SudachiTra

本プロジェクトでは、国立国語研究所に家庭用の B フレッツ回線を敷設し、当該回線を用いてクロールを実施した。各国の国立図書館が広範なウェブページ保存を目指しているのに対し、本プロジェクトではテキストデータを分析対象として限定していたため、収集対象を.html ファイルおよび.txt ファイルに限定した。

クロールの運用に際しては、robots.txt やメタタグなど、ロボット排除プロトコルに基づくサイト運営者の指示を遵守した。また、クロール開始の 1 か月前より、問い合わせ窓口を設置し、クローラ運用に関する問い合わせに対応する体制を整備した。

クロールの設定検討は段階的に実施された。2012 年 7 月には約 100 万 URL 規模の第一次収集テストを、2012 年 8 月から 9 月にかけては約 1000 万 URL 規模の第二次収集テストを行い、それぞれの結果を踏まえて設定を最適化した。その結果、週次の収集量を約 1000 万 URL に設定し、3 か月ごとに約 1 億 URL の収集を目標とする運用方針を確立した。

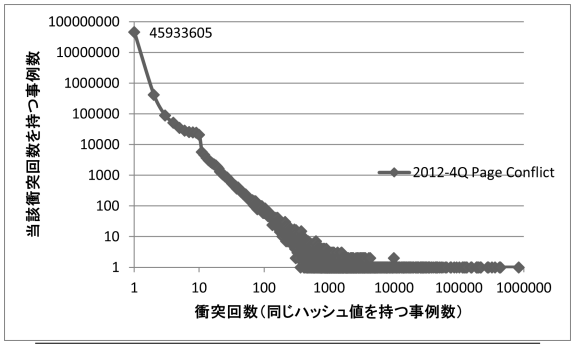
本収集（第一期）は、2012 年第 4 四半期（2012-4Q）から開始された。具体的な運用では、1000 万 URL をクロール可能なインスタンスを 2 つ準備し、これらを 2 週間ごとに実行する体制を採用した。



2012-4Q		
リンク先（のべ）	6,905,806,383	(69 億)
リンク先（異なり）	892,135,930	(8.9 億)

図 1　2012-4Q に含まれる URL

固定した 1 億 URL を 4 回にわたりクロールし、1 年間のクロール終了時点で得られた URL の中から、各クロールで 1 回しかリンクされていない URL を基に 1 億 URL をサンプリングした。この手法は、稀少な言語表現を可能な限り広範に収集するというデータ収集の目的に基づいて採用されたものである。図 1 は、2012 年第 4 四半期（2012-4Q）に収集した URL の被リンク回数を示している。リンク回数が多い URL は、大手ウェブサイトのトップページであることが多かった。本プロジェクトでは、グラフの左上に位置する「被リンク回数が 1」の URL を中心に、次回クロールの対象とする URL をサンプリングした。また、4 回のクロールで収集した同一 URL の中で、内容が毎回変化していた URL も次回クロールの対象として選定し、多様性のあるデータ収集を図った。



2012-4Q	
クロールしたページ数	61,668,805
（内）内容重複なしページ数	45,933,605

図 2　2012-4Q に含まれるページの重複

クロールしたページには、異なる URL であっても内容が重複しているケースが確認された。図 2 は、2012 年第 4 四半期（2012-4Q）に収集したページの内容重複状況を示している。内容が他の URL と重複していないページは 45,933,605 ページであった。一方、重複しているページには、ブログ記事などの同一内容を異なる URL で参照したものやソフト 404 などが含まれていた。

[2012-4Q〜2013-3Q] は最初の 1 億 URL を対象と

表 2　2012 年第 4 四半期から 2013 年第 3 四半期の収集ページ数

	2012-4Q	2013-1Q	2013-2Q	2013-3Q
ページ数　（1 期）	61,668,805	58,844,092	61,479,268	57,892,917
内容の重複なしページ数	45,933,605	42,932,982	45,111,527	42,192,931
4 期通しての統計				
総異なり URL 数（4 期）	64,539,233			
(内) 内容の重複ありページ数	27,604,915			
(内) 内容の重複なしページ数	36,934,706			

表 3　2012 年第 4 四半期から 2013 年第 3 四半期の収集リンク数

	2012-4Q	2013-1Q	2013-2Q	2013-3Q
リンク先（のべ）	6,905,805,383	6,610,763,700	7,064,611,259	7,222,958,033
リンク先（異なり）	892,135,930	843,166,672	865,694,816	855,684,918
4 期通しての統計				
リンク先（異なり）	1,642,699,579			

し、[2013-4Q〜2014-3Q] はその最初の 1 億 URL に含まれるリンクからサンプリングされた 2 番目の 1 億 URL を対象としている。さらに、[2014-4Q〜2015-2Q] では、2 番目の 1 億 URL からサンプリングされた 3 番目の 1 億 URL が対象となっている。

表 2 に収集したページ数の統計量を示す。1 億 URL を収集したものの、robots.txt の順守や各種 HTTP エラーにより、実際にページとして収集できたものは約 6 割にとどまった。重複検出は、URL ごとに各ページのハッシュ値を計算し、同一性を確認する方法で行った。各期において、内容が重複していない（異なり）ページ数は約 4000 万ページ強となった。4 期を通じて、総異なり URL 数は約 6400 万 URL であり、1 億 URL には達しなかった。また、4 期中 2 期以上で収集されたページのうち、内容に重複があるページは約 4 割の 2700 万ページであり、反対に内容に重複がないページは 3600 万ページとなった。

表 3 に 2012 年第 4 四半期（2012-4Q）〜2013 年第 3 四半期（2013-3Q）に収集したリンク数を示す。おおよそ 6000 万 URL の収集に対して、のべ 70 億 URL 前後、異なり URL は約 9 億弱が収集できている。4 期を通じた集計によるリンク先数は異なり 16 億 URL に達しており、これにより、1 年間を通して同じ URL を 4 期にわたり収集することによって、1 期のみのクロールと比べてリンク数が約 1.8 倍（8.5 億〜8.9 億→ 16 億 URL）に増加していることが確認できる。

表 4 に、当時の技術で組織化したデータの基礎統計量を示す。Heritrix は収集した Web ページを圧縮して約 1GB サイズの WARC データに分割して出力する。展開するとそのサイズは約 3 倍となるため、表中の収集 WARC ファイル数に 3GB を掛けた値が、収集した Web ページの容量（メタデータを含む）として概算できる。URL 数は前節で述べた収集 URL 数を基にしている。正規化処理は nwc-toolkit によって行われ、正規化処理の際には文抽出をせずに形態素解析（MeCab/IPADIC）を行った結果、各期における形態素数はのべ約 620〜647 億となった。日本語らしい文を抽出すると、形態素数は各期で約 300 億強となり、これはおおよそ半分の形態素が日本語の文中におけるものではなく、排除されたことを示している。抽出された文数は各期のべ数で約 25 億文前後であり、文単位で同一性を認定した場合、文の異なり数は各期で約 10 億文となった。

4　whole-NWJC:『国語研日本語ウェブコーパス』全データ

2015 年から 2024 年 1 月にかけて公開・共有していたデータは、文単位で異なりを取ったものであった。しかし、大規模言語モデルの訓練には文脈を持つテキストデータが必要であるため、2012-4Q〜2015-2Q の『国語研日本語ウェブコーパス』全データ（whole-NWJC）を共有することとした。

共有するデータは warc.gz 形式で提供され、正規化、日本語文抽出、形態素解析などの処理は行っていないため、利用者自身で処理を行う必要がある。

表 5 に、whole-NWJC の総データ数を示す。2014-1Q の収集量が少ないのは、プロジェクト運営時のストレージ制約によるものである。また、2014-4Q のデータを外部に共有することが決まったため、2015-2Q の途中でクロールを停止した。

表 4　2012 年第 4 四半期から 2013 年第 3 四半期の収集形態素数・文数

	2012-4Q	2013-1Q	2013-2Q	2013-3Q
収集 WARC ファイル	814	870	910	905
URL 数	61,668,805	58,844,092	61,479,268	57,892,917
形態素数（文抽出前）	64,714,650,129	62,077,520,745	63,414,252,638	65,736,027,334
形態素数（文抽出後）	33,767,409,441	32,651,138,004	33,073,991,355	30,923,912,566
文数 （のべ）	2,678,315,774	2,600,122,908	2,659,617,620	2,478,309,312
文数 （異なり）	1,097,011,506	1,048,772,913	1,063,649,324	1,007,771,383

表 5　whole-NWJC: 総データ量

[1st 1 億 URL]	2012-4Q	2013-1Q	2013-2Q	2013-3Q
warc ファイル数	910	878	910	906
ファイルサイズ（圧縮）	842GB	813GB	844GB	838GB
[2nd 1 億 URL]	2013-4Q	2014-1Q	2014-2Q	2014-3Q
warc ファイル数	998	437	1021	608
ファイルサイズ（圧縮）	928GB	407GB	952GB	562GB
[3rd 1 億 URL]	2014-4Q	2015-1Q	2015-2Q	
warc ファイル数	907	874	20	
ファイルサイズ（圧縮）	845GB	812GB	19GB	

5　おわりに

本稿では、2024 年度に新たに共有を開始した『国語研日本語ウェブコーパス』全データ（whole-NWJC）について解説した。このデータは、国内の研究機関の方は国立国語研究所の共同利用型共同利用 (C) [5] に申請することにより利用できる。一般企業の方は、国立国語研究所との共同研究契約を結ぶことで利用可能である。

謝辞

本研究は国立国語研究所「超大規模コーパス」プロジェクト (2011-2015) によるものです。また chiVe および chiTra は、ワークス徳島人工知能 NLP 研究所と国立国語研究所の共同研究によるものです。

参考文献

[1] Masayuki Asahara. NWJC2Vec: Word embedding dataset from 'NINJAL Web Japanese Corpus'. **Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication**, Vol. 24, No. 2, pp. 7–25, 2018.

[2] 新納浩幸, 浅原正幸, 古宮嘉那子, 佐々木稔. nwjc2vec: 国語研日本語ウェブコーパスから構築した単語の分散表現データ. 自然言語処理, Vol. 24, No. 5, pp. 705–720, 2017.

[3] 真鍋陽俊, 岡照晃, 海川祥毅, 髙岡一馬, 内田佳孝, 浅原正幸. 複数粒度の分割結果に基づく日本語単語分散表現. 言語処理学会第 25 回年次大会 (NLP2019), pp. NLP2019–P8–5. 言語処理学会, 2019.

[4] 河村宗一郎, 久本空海, 真鍋陽俊, 髙岡一馬, 内田佳孝, 岡照晃, 浅原正幸. chive 2.0: Sudachi と nwjc を用いた実用的な日本語単語ベクトルの実現へ向けて. 言語処理学会第 26 回年次大会 (NLP2020), pp. NLP2020–P6–16. 言語処理学会, 2020.

[5] 久本空海, 山村崇, 勝田哲弘, 竹林佑斗, 髙岡一馬, 内田佳孝, 岡照晃, 浅原正幸. chive: 製品利用可能な日本語単語ベクトル資源の実現へ向けて. 第 16 回テキストアナリティクス・シンポジウム, pp. IEICE–NLC2020–9. 電子情報通信学会, 2020.

[6] 浅原正幸, 西内沙恵, 加藤祥. NWJC-BERT: 多義語に対するヒトと文脈化単語埋め込みの類似性判断の対照分析. 言語処理学会第 26 回年次大会発表論文集, pp. 961–964, 2020.

[7] 勝田哲弘, 林政義, 山村崇, Tolmachev Arseny, 髙岡一馬, 内田佳孝, 浅原正幸. 単語正規化による表記ゆれに頑健な bert モデルの構築. 言語処理学会第 28 回年次大会 (NLP2022). 言語処理学会, 2022.

[8] Masayuki Asahara and Kikuo Maekawa. Design of a Web-scale Japanese Corpus. In **Proceedings of the Conference of the Pacific Association for Computational Linguistics (PACLING-2013)**, 2013.

[9] 浅原正幸, 今田水穂, 保田祥, 小西光, 前川喜久雄. Web を母集団とした超大規模コーパスの開発 収集と組織化. 国立国語研究所論集, No. 7, 5 2014.

[10] Masayuki Asahara, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi. Archiving and Analysing Techniques of the Ultra-large-scale Web-based Corpus Project of NINJAL, Japan. **Alexandria**, Vol. 26, No. 1-2, pp. 129–148, 2014.

[11] Masayuki Asahara, Kazuya Kawahara, Yuya Takei, Hideto Masuoka, Yasuko Ohba, Yuki Torii, Toru Morii, Yuki Tanaka, Kikuo Maekawa, Sachi Kato, and Hikari Konishi. 'BonTen' - Corpus Concordance System for 'NINJAL Web Japanese Corpus'. In **Proceedings of COLING-2016 Demo Session**, 2016.

[12] 浅原正幸, 河原一哉, 大場寧子, 前川喜久雄. 『国語研日本語ウェブコーパス』とその検索系『梵天』. 情報処理学会論文誌, Vol. 59, No. 2, pp. 299–306, 2018.

5）https://www.ninjal.ac.jp/research/cfp/jupc/