whole-NWJC: 『国語研日本語ウェブコーパス』全データ

浅原正幸 [1]
[1] 国立国語研究所　次世代言語科学研究センター
masayu-a@ninjal.ac.jp

概要

本稿では、whole-NWJC:『国語研日本語ウェブコーパス』(NINJAL Web Japanese Corpus: NWJC) の全データについて解説する。同データは国立国語研究所と共同研究を申請することで利用できる。

キーワード：ウェブコーパス, ウェブクローラ

1　はじめに

本稿では、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」で整備した『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 について解説する。『国語研日本語ウェブコーパス』は 2014 年第 4 四半期 (2014-4Q) を検索系「梵天」「中納言」で共有した[1)]ほか、語彙表・n-gram データの形式で共有してきた。今回、大規模言語モデルに資するデータとして、2012 年第 4 四半期から 2015 年第 2 四半期に収集した WARC 形式のデータを共有する。以下では共有するデータの概要について解説する。

2　既共有データ

設計時の情報を文献 [1, 2, 3] にまとめた。

当初のプロジェクトでは、検索系「梵天」[4, 5] により 2014-4Q のデータ（収集 URL 数 83,992,556、文数（のべ数） 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）を公開していたが、2021 年 12 月 24 日にサーバ維持費用の制約があり、共有を停止した。また、検索系「中納言」上で 2014-4Q の一部データ（国語研短単位数 86,277,772：NWJC-2014-4Q の 0.33%）を公開していたが、2024 年 2 月 29 日に共有を停止した。

統計情報として、語彙表・中納言搭載データ語彙表・n-gram データを公開した。

また、単語埋め込みモデルとして NWJC2vec [6, 7] と chiVe [8, 9, 10] を公開した。さらに訓練済みの BERT モデルとして、NWJC-BERT [11]（語彙素に基づき文単位で訓練）と chiTra [12]（文章単位で訓練）を公開した。

3　データ収集の概要

『国語研日本語ウェブコーパス』(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て構築したコーパスである。当初は 100 億語規模のコーパスを構築し、外部に検索系を介して共有することを目標としていた。

ページ収集は Heritrix クローラ [2)] を用いた 遠隔採取 (remote harvesting) による。Heritrix クローラは Internet Archive が開発したウェブアーカイブのための Web クローラである。各国の国立図書館等が国内外の Web ページを保存するために用いられている。日本国内でも国立国会図書館がインターネット資料収集保存事業 [3)] において用いられている。Heritrix は ISO 28500 で規定される WARC ファイル形式でウェブアーカイブを保存する。これは複数のウェブ資料をその HTTP ヘッダやレスポンスなどの情報とともに、単一のファイルに格納したものである。Python のライブラリ [4)] を用いることでデータ処理を行える。

我々は国立国語研究所に家庭用の B フレッツ回線を引き、その回線からクロールを進めた。各国の国立図書館と異なり、本プロジェクトではテキストのみが分析対象であったために、.html ファイル・.txt ファイルに限定して収集した。クローラの運用においては、robots.txt およびメタタグなどのロボット排除プロトコルを確認し、サイト運営者側のクローラプログラムへの支持を遵守した。さらにクローラの試験運用 1 カ月前より、クローラに関する問い合

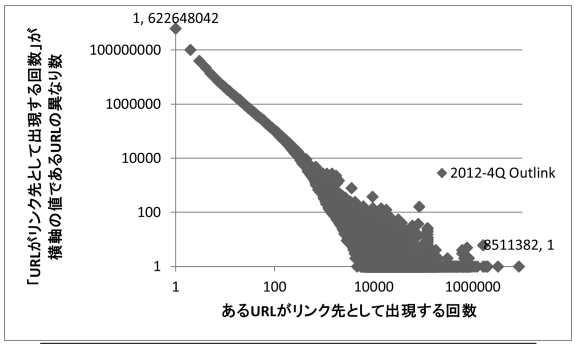
1）「梵天」は 2021 年 12 月 24 日に共有停止、「中納言」2024 年 2 月 29 日に共有停止。

2）https://github.com/internetarchive/heritrix3
3）https://warp.da.ndl.go.jp/
4）https://github.com/internetarchive/warc

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 [6, 7]	https://www.gsk.or.jp/catalog/gsk2020-d
chiVe: Sudachi と NWJC による日本語単語ベクトル [8, 9, 10]	https://github.com/WorksApplications/chiVe
「国語研日本語ウェブコーパス」NWJC-BERT [11]	https://www.gsk.or.jp/catalog/gsk2020-e
chiTra: Sudachi と NWJC による Transformer モデル [12]	https://github.com/WorksApplications/SudachiTra

表 1　『国語研日本語ウェブコーパス』既共有データ

わせ窓口を設置した。2012 年 7 月に 100 万 URL 規模の第一次収集テスト，2012 年 8〜9 月に 1000 万 URL 規模の第二次収集テストを繰り返し行い，クローラの設定を検討した結果，週次の収集量を 1000 万 URL 程度とし，3 か月ごとに 1 億 URL 規模の収集を行うことにした。2012 年第 4 四半期（2012-4Q）から本収集（第一期）を開始した。具体的には 1000 万 URL クロールするインスタンスを 2 つ準備し、2 週間ごとに実行する運用を行った。

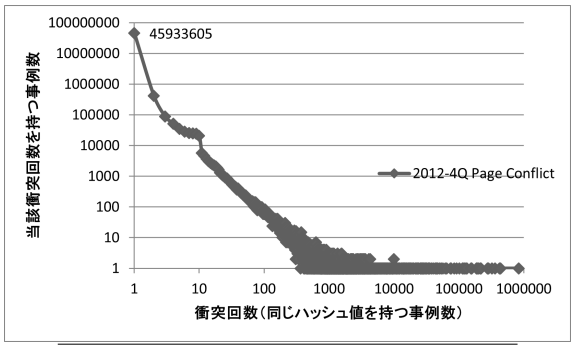


2012-4Q		
リンク先（のべ）	6,905,806,383	(69 億)
リンク先（異なり）	892,135,930	(8.9 億)

図 1　2012-4Q に含まれる URL

固定した 1 億 URL を 4 回クロールし、1 年クロールが終了した時点で得られた URL のうち、各クロールで 1 回しかリンクされていないものから 1 億 URL をサンプリングした。これは、データ収集の目的として、稀な言語表現をできるだけ浅く広く収集することを目的としたためである。図 1 に 2012-4Q (2012 年第 4 四半期) に収集した URL の被リンク回数を示す。リンク回数が多いものは、大手サイトのトップページであったり robots.txt であった。グラフで左上に示されている「被リンク回数」が 1 のものを中心に次回クロールすべき URL をサンプリングした。そのほか、4 期にクロールした同一 URL のうち内容が毎回変わっていた URL からも、次回クロールすべき URL をサンプリングした。

クロールしたページは重複が確認された。図 2 に 2012-4Q に収集したページの重複を示す。他の URL



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

と内容重複がないページは 45,933,605 ページであった。重複されているページは、ブログなどの同一ページを異なる URL で示したもののほか、robots.txt やソフト 404 であった。

[2012-4Q〜2013-3Q] が最初の 1 億 URL、[2013-4Q〜2014-3Q] が最初の 1 億 URL に含まれるリンクからサンプリングした 2 番目の 1 億 URL、[2014-4Q〜2015-2Q] が 2 番目の 1 億 URL からサンプリングした 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 を収集しても robots.txt の順守や各種 HTTP エラーにより，ページとして収集できたものが約六割にすぎない。重複検出は URL 毎に各ページのハッシュ値を計算し同一性を認定する。各期において内容の重複なし（異なり）ページ数は 4000 万強になる。4 期通しての総異なり URL 数は約 6400 万 URL と 1 億 URL に至らない。4 期中 2 期以上収集できたページ数の内，内容の重複があるページ数は約四割の 2700 万ページ，反対に内容の重複がないページ数は 3600 万ページになる。

表 3 に 2012 年第 4 四半期（2012-4Q）〜2013 年第 3 四半期（2013-3Q）の収集リンク数を示す。おおよそ 6000 万 URL の収集に対し，のべ 70 億前後，異なり 9 億弱の URL が収集できている。4 期を通した集計によるリンク先数が異なり 16 億 URL であるこ

	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 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
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			

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
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 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

とから 1 年間通して同じ URL を 4 期収集することにより 1 期のみクロールするのみ比べてリンクが約 1.8 倍（8.5 億〜8.9 億→ 16 億 URL）に増えていることがわかる。

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 Web ページを圧縮約 1GB サイズの WARC データに分割して出力する。展開すると約 3 倍程度になるため，表中の収集 WARC ファイル数に 3GB をかけた値が収集 Web ページ容量（メタデータを含む）と概算することができる。URL 数は前節の収集における URL 数である。正規化処理は nwc-toolkit による。正規化処理の際に文抽出なしに形態素解析（MeCab/IPADIC）を行うと各期のべ約 620〜647 億形態素になる。日本語らしい文の抽出を行うと形態素数は各期約 300 億強になることから大体半分の形態素が日本語の文中の形態素ではないとして排除されている。抽出された文数はのべ数で各期 25 億文前後，文単位の同一性を認定すると文の異なり数は各期 10 億文になる。

4　whole-NWJC

2015 年から 2024 年 1 月にかけて公開していたデータは、文単位の異なりを取ったものであった。大規模言語モデルの訓練においては、文脈を持つテキストデータが必要となる。そこで、2012-4Q〜2015-2Q の NWJC データ (whole-NWJC) を共有する。共有するデータは warc.gz 形式で、正規化・日本語文抽出・形態素解析などは行っていないために、利用時には各自処理する必要がある。

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の制約による。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。

5　おわりに

本稿では、『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。同データは、国立国語研究所 共同利用型共同利用 (C) [5] に申請するか、共同研究契約を結ぶことで利用することができる。

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

[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	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	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	845GB	812GB	19GB	

表 5　whole-NWJC: 総データ量

謝辞

本研究は国立国語研究所コーパス開発センター「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 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 河村宗一郎, 久本空海, 真鍋陽俊, 高岡一馬, 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.

[10] 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.

[11] 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.

[12] 勝田哲弘, 林政義, 山村崇, Tolmachev Arseny, 高岡一馬, 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.