Web を母集団とした超大規模コーパスの設計

浅原 正幸 †　　前川 喜久雄 ‡
国立国語研究所　コーパス開発センター †
国立国語研究所　言語資源研究系/コーパス開発センター ‡
masayu-a -at- ninjal.ac.jp, kikuo -at- ninjal.ac.jp

1　はじめに

国立国語研究所では 2006 年-2010 年の期間に 1 億語規模の書き言葉コーパス『現代日本語書き言葉均衡コーパス』(BCCWJ)[1, 2] を構築し、2011 年より一般公開している。BCCWJ は種々の母集団に沿った無作為抽出を実施することによって、高度な代表性を備えた均衡コーパスとなっている。しかし、その規模は、現代のコーパス言語学の趨勢からすれば十分とはいいがたく、生起頻度の低い言語現象の被覆に問題がある。そのためより大規模な日本語コーパスの構築が望まれている。この問題を解消するため、国立国語研究所では 2011 年から 6 か年の期間で、Web を母集団とした 100 億語規模の超大規模コーパスを構築する計画に着手した。本発表では、超大規模コーパスをどのようにして構築するか、どのような情報を付与するか、どのような検索環境を提供するのかなど、超大規模コーパスの設計の概要について解説する。

2　超大規模コーパスの概要設計

本節では構築する超大規模コーパスの概要設計 (表 1) について、収集・構造化・利活用・永続保存の四つの観点から解説する。一つ目は Web テキストをクロールする指針について説明する。二つ目は Web コーパスとして利用可能にするための正規化技術・形態素解析・係り受け解析・レジスタ推定について説明する。三つ目は Web コーパスの利活用として検索サービス・語彙表/n-gram データ整備・言語解析技術での利用などについて述べる。四つ目は Web アーカイブとして永続保存する技術について述べる。

表 1: 超大規模コーパスの設計：まとめ

収集		
	クローラ	
		Heritrix 3.1 系
構造化		
	正規化技術	
	形態素解析	
		JUMAN(益岡田窪品詞体系)
		MeCab/UniDic (国語研短単位, UniDic 品詞体系)
		CRF++ (国語研長単位, UniDic 品詞体系)
		教師なし形態素解析 (単語分かち書きのみ)
	係り受け解析	
		CaboCha/京都大学テキストコーパス
		CaboCha/BCCWJ
		教師なし係り受け解析
	レジスタ分析	
		BCCWJ メタデータ相当情報推定 (教師あり機械学習)
		splog 等判定 (教師あり機械学習)
		クラスタリングによる文体論的分析
利活用		
	検索サービス	
		文字列検索 (＋レジスタによるファセット分析)
		品詞検索 (中納言相当)
		係り受け検索 (ChaKi Dependency Search 相当)
	語彙表・n-gram データ	
		語彙表 (出現形, 形態論情報を含む)
		n-gram データ (基本形, 形態論情報を含まない)
		係り受け部分木 (基本形, 形態論情報を含まない)
	言語解析器	
		UniDic 未登録語調査
		頻度・共起情報を用いた言語解析器の改善
永続保存		
	ファイル形式	
		WARC 形式 (ISO-28500)
	情報アクセス	
		Open Source Wayback (ハーベスト)
		NutchWAX (検索)
	キュレーション	
		Web Curator Tool

2.1　収集

Web コーパスを構築するためにデータを収集する方法として、検索エンジンが提供する Web API を利用する方法と、自分でクローラを運用して収集する方法がある。本研究では後者のクローラを用いる手法を採用する。Web テキストの収集はクローラ Heritrix [1]を利用する。Heritrix クローラは、wayback machine と呼ばれる Web アーカイブに実績を持つ米国 Internet Archive が中心となり開発しているクローラプログラ

[1] webarchive.jira.com/wiki/display/Heritrix/Heritrix

ムである。各国国立図書館が Web アーカイブを構築するために利用しており、日本では国立国会図書館がインターネット資料収集保存事業において利用している。アーカイブの保存形式は、後述する Web アーカイブの標準化 WARC 形式が選択できる。

各国国立図書館で運用するクローラは画像ファイル・音声ファイル・動画ファイルも含めたバルク収集ができることが重要であるが、本研究においてはテキストデータの収集が主な目的であるために、.html ファイル・.txt ファイル・.xml ファイルに限定して収集している。

1 億 URL ほどをシード URL リストとして、年に 4 回のペースで定点観測的に Web テキストとリンク-被リンク構造の収集を行う。

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 5000 万 URL 規模の第二次収集テストを行い、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 年第四四半期から本収集（第一期）開始し、2013 年 1 月現在、本収集（第二期）を行っている。

収集対象は基本的に日本語である。日本語であれば splog であろうと機械翻訳結果であろうと収集を行い保存する。外国語などのデータは後述するレジスタ推定などにより定期収集対象から除外するが、収集したものを削除することは行わない。

今後一年かけて同様の本収集を行い、 URL の更新頻度推定などを行う。二年目以降は更新されない URL を収集範囲から外したうえで、新しい URL を収集範囲として含め、収集範囲の拡充を行う。

2.2 構造化

Web テキストは収集しただけではそのままコーパスとして用をなさない。以下では、HTML タグ排除や文字の統制などの**正規化**、言語解析としての**形態素解析**、**係り受け解析**、コーパスとしての母集団を規定するための基礎情報となる **レジスタ推定** について説明する。

正規化：収集した Web テキストは、nwc-toolkit [2]を用いて、Google「Web 日本語 N グラム第 1 版」[3]に準じた正規化を行う。具体的には文字コード判別・変換、HTML の解析、テキストの抽出、Unicode 正規化、句点などによる文分割、文字数や文字種によるフィルタリングなどを行う。正規化を行った Web テキストに対して、形態素解析・係り受け解析・レジスタ分析を行い、構造化を行う。

形態素解析：形態素解析手法として、JUMAN [4] による益岡田窪品詞体系に基づく解析、MeCab[5]/UniDic[6] による国語研短単位解析、汎用チャンカー CRF++ [7]による国語研長単位解析、ベイズ階層言語モデルによる教師なし形態素解析 [5] の四つを検討している。

係り受け解析：係り受け解析手法として、京都大学テキストコーパス[8]の基準に基づいて学習した CaboCha [9]による解析・BCCWJ 基準 [4] に基づいて学習した CaboCha ・ベイズ階層言語モデルによる教師なし係り受け解析器（開発予定）の三つを検討している。

レジスタ推定：言語学の観点からすると、Web コーパスの可用性を下げる大きな要因のひとつは、収集されたテキストがどのような目的で書かれているかというレジスタ情報の欠落である。そのため本コーパスでは、サンプルのレジスタ推定を実施する。収集の時点では、シード URL からリンク構造をたどることによりクロールするため、自然言語コーパスとして均衡性・代表性を持たせた母集団を規定することが困難である。あらかじめ文書分類的な手法を用いて適切な部分サンプル集合をレジスタとして規定することにより、この問題を緩和する。具体的には、外国語・splog・機械翻訳や機械生成されたテキストで非文法的なものを排除するための分類（(半) 教師あり機械学習）、BCCWJ に付与された各種メタデータ・ファイル単位アノテーションを推定するための分類（(半) 教師あり機械学習）、クラスタリングに基づく分類（教師なし機械学習）などを検討している。教師あり機械学習においては、多クラスの Transductive SVM [10]による境界事例分析と、ランダムフォレスト法やブースティング法[11]による有効特徴量分析を行い、クラスタリングによる分類については得られたクラスタに対して言語学（文体論）的な見地からの分析を行う。

[2] code.google.com/p/nwc-toolkit/
[3] www.gsk.or.jp/catalog/GSK2007-C/GSK2007C_README.utf8.txt
[4] nlp.ist.i.kyoto-u.ac.jp/nl-resource/juman/juman-7.0.tar.bz
[5] mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[6] sourceforge.jp/projects/unidic/
[7] crfpp.googlecode.com/svn/trunk/doc/index.html
[8] nlp.ist.i.kyoto-u.ac.jp/nl-resource/corpus/KyotoCorpus4.0.tar.gz
[9] code.google.com/p/cabocha/
[10] vikas.sindhwani.org/svmlin.html
[11] chasen.org/~taku/software/bact/

2.3 利活用

構造化されたコーパスとして利活用していくうえで必要な環境整備について、**検索サービス** と **語彙表・n-gram データ** について説明する。また利活用の事例として想定している**言語解析技術への利用** についても述べる。

検索サービス：構築したコーパスについては、外部向けの検索サービスを提供することを想定している。レジスタに基づいた絞込を可能にする高速文字列検索機能、コーパス検索アプリケーション「中納言」[12]のような品詞に基づいた検索機能、コーパス開発環境「ChaKi」[13]の Dependency Search のような係り受け構造に基づいた検索機能を、100 億語規模で現実的に動作する機能に絞って提供する予定である。

語彙表・n-gram データ：3 カ月おきにクロールするデータに対して構造化を行ったうえで、語彙表 (1-gram 頻度情報; 形態論情報を含む; 出現形に基づく)・系列上の n-gram 頻度情報 ($n \geq 1$; 形態論情報を含まない; 基本形に基づく)・係り受け木上の部分木頻度情報を収集データ全体と推定したレジスタごとに取得する。単位としては、益岡田窪文法による JUMAN 単位、国語研短単位、国語研長単位の三つとし、係り受け木については京都大学テキストコーパス基準と BCCWJ 基準の二つを想定している。n-gram データの構築には FREQT[14]を利用する。また、別処理により、HTML タグの頻度情報・リンク-被リンク関係・同一コンテンツ関係など Web テキスト特有の情報を取得し保持する。可能であれば、レジスタ推定時にこれらの情報を活用する。

言語解析技術への利用：得られたコーパスを用いた言語解析技術の向上手法について検討を行う。形態素解析においては、教師なし形態素解析技術や未知語処理技術により得られた UniDic 未登録語について、人手で形態論情報を付与することにより辞書の拡充を行う。n-gram 頻度情報や部分木頻度情報を用いた各種言語解析技術の性能改善手法について検討を行う。

2.4 永続保存

収集したデータは言語の経年変化を分析するための基礎データとするために永続保存する。IIPC(International Internet Preservation Consortium)[15]における各国国立図書館の活動動向を見ながら、保存のための構造化を行う。具体的には Heritrix で収集されたデータは、Web アーカイブの保存形式の国際標準 WARC 形式[16]で保存する。WARC ファイルは Internet Archive が公開している Wayback Machine [17]と同じ機能を持つハーベストソフトウェア Open Source Wayback[18] と、情報検索システム NutchWAX (Nutch Web Archive eXtension) [19]により構造化し、Web アーカイブとしての情報アクセスを可能にする。また、選択的な Web クロールを可能にするためのキュレーションツール WCT (Web Curator Tool)[20]の技術調査を行う。最後に、長期保存可能な記憶媒体を確保し、収集し構造化したデータの保存に努める。

3 おわりに

表 2 に現状の工程表を示す。2011 年度後半に計画立案を行った。2012 年度は主に収集技術・テキストの正規化技術・形態素解析技術・文字列検索技術・保存技術の調査を行った。収集技術に関しては実際にクローラの運用テストを行いながら運用規則の策定を行い、現在クローラの本運用を開始している。今後定期的に運用規則を見直しながら収集作業をすすめる。2013 年度は、テキストの正規化技術・形態素解析関連技術を運用レベルにあげ、文字列検索技術の調達を開始する。係り受け解析技術の既存技術については調査を行うとともに年度末までに運用レベルにする。技術調査としてはレジスタ分析技術と品詞・係り受け構造に基づく検索技術を対象とする。2014 年度以降、細部については修正の可能性もあるが、大方はこの工程表に準じて構築をすすめる予定である。

謝辞

本研究は国立国語研究所コーパス開発センターの「超大規模コーパス構築プロジェクト」によるものである。

本研究を行うにあたり、情報通信研究機構ユニバーサルコミュニケーション研究所の諸氏および統計数理研究所の持橋大地氏よりさまざまな技術指導をいただいた。国立国語研究所コーパス開発センターの諸氏から設計時点での有益なコメントをいただいた。ここに記して謝意を表す。

[12] chunagon.ninjal.ac.jp
[13] sourceforge.jp/projects/chaki/releases/
[14] chasen.org/~taku/software/freqt/
[15] netpreserve.org/
[16] ISO 28500:2009, Information and documentation – WARC file format
[17] archive.org/web/web.php
[18] archive-access.sourceforge.net/projects/wayback/
[19] archive-access.sourceforge.net/projects/nutch/
[20] webcurator.sourceforge.net/

表 2: 超大規模コーパスプロジェクト：工程表

年	2012				2013				2014				2015				2016
四半期	1Q	2Q	3Q	4Q	1Q	2Q	3Q	4Q	1Q	2Q	3Q	4Q	1Q	2Q	3Q	4Q	1Q
準備	⇒ 計画立案 ⇒ 機材調達 (初回)		⇒ (2 回目)				⇒ (3 回目)				⇒ (4 回目)						
収集	(クローラ関連) ⇒ クローラ運用テスト ⇒ クローラ本運用開始 ⇒ 運用規則策定				⇒ 運用規則見直し (初回)				⇒ 運用規則見直し (2 回目)				⇒ 運用規則見直し (3 回目)				
構造化	(正規化) ⇒ 正規化技術調査 ⇒ 正規化技術運用開始																
	(形態素解析) ⇒ 形態素解析技術調査（既存技術） ⇒ 形態素解析運用開始（既存技術）																
	(係り受け解析) ⇒ 係り受け解析技術調査（既存技術） ⇒ 係り受け解析運用開始（既存技術） ⇒ 係り受け解析技術調査・開発（新規技術）																
	(レジスタ分析) ⇒ BCCWJ メタデータ関連調査 ⇒ splog 検出技術調査 ⇒ クラスタリングによる文体論的分析技術調査 ⇒ 実装・並列化 ⇒ レジスタ分析技術運用開始																
利活用	(検索サービス) ⇒ 文字列検索技術調査 ⇒ 文字列検索技術調達開始 ⇒ 品詞検索技術調査 ⇒ 係り受け検索技術調査								⇒ 品詞検索技術調達開始 ⇒ 係り受け検索技術調達開始							⇒ 内部公開 ⇒ 外部公開	
	(語彙表・n-gram データ) ⇒ 語彙表作成開始 ⇒ n-gram データ作成開始 ⇒ 係り受け部分木データ作成開始																
	(言語解析器) ⇒ 未登録語調査 (初回)								⇒ (2 回目)				⇒ (3 回目) ⇒ 言語解析器の改善				
保存	⇒ 技術調査 ⇒ Open Source Wayback 運用開始 ⇒ NutchWAX 運用開始 ⇒ 保存媒体の確保																

参考文献

[1] 前川喜久雄. コーパス日本語学の可能性—大規模均衡コーパスがもたらすもの—. 日本語科学, 22, pp. 13–28, 11 2007.

[2] 前川喜久雄, 山崎誠. 『現代日本語書き言葉均衡コーパス』. 国文学解釈と鑑賞, 932(74 巻 1 号), pp. 15–25, 12 2008.

[3] 浅原正幸, 前川喜久雄. Web を母集団とした超大規模コーパスの設計. 第 3 回コーパス日本語学ワークショップ論文集, 2013.

[4] 浅原正幸, 松本裕治. 『現代日本語書き言葉均衡コーパス』に対する係り受け・並列構造アノテーション. 第 19 回言語処理学会年次大会 (NLP2013), 2013.

[5] 持橋大地, 山田武士, 上田修功. ベイズ階層言語モデルによる教師なし形態素解析. 情報処理学会研究報告, No. 2009-NL-190, 2009.