E-004

Web からの知識獲得格差実態把握のための言語分布調査モジュール開発に向けた基礎調査

Basic research for development of language distribution survey module for grasping actual knowledge acquisition gap from the Web

武田 大河 † 中平 勝子 † 北島 宗雄 †
Taiga Takeda Katsuko T. Nakahira Muneo Kitajima

1 はじめに

世界に現存している言語は Ethnologue * によれば 6900 超であるが，コンピュータ表現可能な言語は，その 10% 程度に過ぎない．この状況は，情報獲得におけるデジタルデバイドと捉えることができ，その状況把握には，多くの言語を判別可能な言語判定エンジンが必要である．これまで，n グラムパタンを用いた言語判定エンジンが主流で，特に多くの言語 (440 言語) に翻訳された文書である世界人権宣言 (UDHR) を教師データとしたエンジンには Chew ら [3] や Vatanen ら [2] のものがある．いづれも，多言語/多 LSE 判定に有益なエンジンとして提供されているが，手法の限界については示されていない．本稿では，Web 空間における情報表示手段の一つである文書データに着目し，文書がもつ言語情報をバイナリ n グラムパタン (BNP) によって解析し，次世代の言語判定エンジン設計に向けた課題抽出を行う．Web 言語工学の主流である LSE(Language, Script, Encoding)[1] を適用し，予め LSE が特定されている文書中の BNP に対し固有 BNP を持つ LSE を特定し，n グラム解析で判別可能な LSE 数，判別不能な LSE 数を特定し，その分布から言語判定エンジン設計に向けた課題抽出を行う．

2 LSE 毎の固有 BNP 分布分析

分析対象のテキストとして，UDHR を選定した．この文書に対し，LSE で分類した N_{LSE} 種の電子テキストデータから BNP を作成した．バイト列の規模は，$n = 3$ が最も言語判定の精度が高いため，本稿では 3-グラムで解析を行った [3].

本稿の解析では，LSE に分類された上位 10 個の BNP を抽出している．これは，n グラム解析する際に LSE ごとで出現数に大きく差があるためである．例として，(L, S, E) = (russian, cyrillic, cyrillic) の BNP の最頻値は 631 であるが，(japanese, japanese, SJIS) では 131 である．このため，出現数で閾値を設けた場合，言語ごとで得られる BNP に差が生じるためである．

LSE 毎の固有 BNP を算出する際， 文字コードは次の様に分類した．

- utf8: 世界標準の文字コードとなっており，web ページ上の文字コードの 90% を占めている．
- latin1: 2010 年では全電子文書に対する latin1 エンコード使用率は 50% を占めていたものの 2017 年では 4% になっている．しかし，古くから存在する web ページでは latin1 を使用しているため，無視できるものではない．
- その他: 日本語の SJIS や中国語の Big-5 など，種類は多いが，latin1 ほど利用範囲は広くないため、今回の調査ではその他として分類することとした．

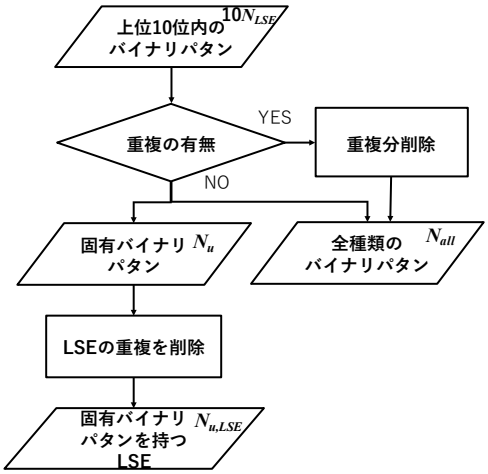

図 1 BNP データ作成フローチャート

分析手順は次の通りである．図 1 にその手順を示す．N_{LSE} の LSE セットからそれぞれ上位 10 個の BNP を抽出する．その総数は $10N_{LSE}$ となる．このデータから，重複のない全種類の BNP データ（図 1 中 N_{all}）を作成する．作成の手順は，BNP が重複の有無で分岐を設け，重複分については重複を除いて追加し，重複していない BNP（図中 N_u）はすべて追加する．

先ほど得られた N_u は，他の LSE に出現する BNP を含んでいる．n グラムによる LSE 判別には他の LSE には出現しない固有の BNP が必要である．固有 BNP データの作成手順は，BNP の重複のない，つまり 1 つの LSE でしか出現しない BNP（図中 $N_{u,LSE}$）を追加することで作成可能である．

N_{all} と N_u は文字コードで分類することで，文字コード毎の固有 BNP 分布を算出することができる．N_u が多く存在する文字コードでは，少ないものと比べ精度の高い LSE 判別が可能となる．逆に N_u が少ない文字コードは固有の n グラムによる LSE 判別は低くなると考えられる．

N_u をもつ LSE は，n グラム解析によって得られる N_u によって判別が可能となる．しかし，LSE の中には N_u を持たないものも存在し，これらは N_u による判別はできない．

N_u を持つ LSE も文字コードで分類することで，文字コード毎の固有 BNP 分布を算出することができる．N_u をもつ LSE

† 長岡技術科学大学

* https://www.ethnologue.com/ethnoblog/m-paul-lewis/ethnologue-launches-subscription-service#.Voav2LaLRdh

表 1　BNP 分布表.

エンコード種別		latin1	utf8	その他
上位 10 位内の BNP 総計	5700			
全種類の BNP	1886	652	836	398
固有 BNP	1121	227	582	312
総 LSE 数	570	177	330	63
固有 BNP を持つ LSE 数	303	104	150	49

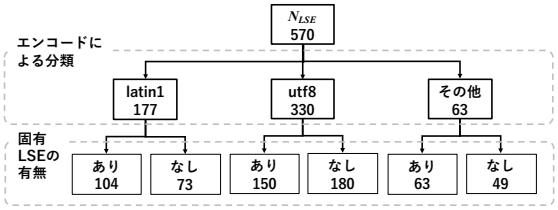

図 2　文字コードごとの固有 BNP をもつ LSE 分布チャート

が多く存在する文字コードでは，少ないものと比べ多くの LSE が N_u による判別が可能となる．逆に N_u をもつ LSE が少ない文字コードでは，固有の n グラムによる判別は難しいと考えられる．

3　分析結果

本項では，N_{LSE} を 570 とした．したがって，上位 10 個の BNP を抽出すれば 5700 個の BNP10N_{LSE} が得られる．このうち図 1 のデータフロー図に従い、N_{all} を算出した結果、1886 個となった．

N_{all} を文字コードにより分類すると latin1 は 652 個，utf8 は 836 個，その他の文字コードでは 398 個存在する．latin1 : $N_{all,latin1} = 652$ utf8 : $N_{all,utf8} = 836$ その他 (other) : $N_{all,other} = 398$

N_{all} から重複を削除したものが N_u である．この固有 BNP はほかの LSE では出現しない BNP であり，固有 BNP が存在する LSE では n グラムによる判別が可能となる．この処理の結果を表 1 に示す.

先ほど得られたデータを基に文字コード事の固有 BNP 分布を算出する導出の計算は、例えば latin1 の場合であれば $N_{u,latin1}/N_{all,latin1}$*100 で求めることができる.

- latin1: 固有の BNP は $N_{u,latin1} = 227$ よって、全体の約 35% が固有の BNP をもつ.
- utf8: 固有の BNP は $N_{u,utf8} = 582$ よって、全体の約 70% が固有の BNP をもつ.
- その他: 固有の BNP は $N_{u,other} = 312$ よって、全体の約 78% が固有の BNP をもつ.

次に LSE ごとの固有 BNP 分布を算出する．全 LSE は 570 存在し，うち latin1 で文字コードされたものは 177，utf8 では 330 あり，その他の文字コードでは 63 存在する．latin1 : $N_{all,LSE,latin1} = 177$ utf8 : $N_{all,LSE,utf8} = 330$ その他 (other) : $N_{all,LSE,other} = 63$

- latin1: 固有の BNP は $N_{u,latin1} = 177$ よって、全体の約 59% の LSE が固有の BNP をもつ.
- utf8: 固有の BNP は $N_{u,utf8} = 330$ よって、全体の約 45% の LSE が固有の BNP をもつ.
- その他: 固有の BNP は $N_{u,other} = 63$ よって、全体の約 78% の LSE が固有の BNP をもつ.

この結果を図 2 に示す.

4　考察

latin1 では、全体の $N_{all,latin1}$ に対する N_{all} の存在比は 35% であり、$N_{LSE,latin1}$ に対する $N_{u,LSE,latin1}$ の存在比は 59% であった．このことから，latin1 で文字コードされている LSE の特徴として，全 BNP のうち固有 BNP は少ないが，LSE ごと均等に固有 BNP が存在していると考えられる．つまり，1 つの LSE に存在する固有 BNP は少ないが，6 割の LSE に固有 BNP が存在するということである．よって、LSE 判別の精度は低い可能性があるものの、約 6 割の LSE が固有 BNP による判別が可能である.

一方，utf8 では、全体の $N_{all,utf8}$ に対する N_{all} の存在比は 70% であり、$N_{LSE,latin1}$ に対する $N_{u,LSE,utf8}$ の存在比は 45% であった．このことから，utf8 で文字コードされた LSE の特徴として，全 BNP のうち多数の固有 BNP をもつが，その分布は言語毎に偏りがあると考えられる．つまり，latin1 に比べ 1 つの LSE には多くの固有 BNP をもつが、その LSE 数は全体の 4 割ほどであることを意味する．よって、LSE 判別の精度は高いものの、固有 BNP による判別が可能な LSE は約 4 割ほどである.

その他の文字コードでは、全体の $N_{all,other}$ に対する N_{all} の存在比は 78% であり、$N_{LSE,latin1}$ に対する $N_{u,LSE,other}$ の存在比は 78% であった．このことから、その他に分類されている言語独自の文字コードでは、多くの固有 BNP が多くの LSE に存在していることを意味する．よって約 8 の LSE は固有 BNP による高精度な LSE 判別が可能である.

なお現在，各 LSE から頻出数が上位 10 位までの BNP を教師データとして Wikipedia のテキストデータを判別するテストを行っている.

5　まとめ

今回行った分析により、文字コードごとの固有 BNP 分布のデータを得ることができた．utf8 では固有 BNP を持つ言語が少ないため，判定には固有 BNP だけでは不十分であること．また latin1 においては固有 BNP のデータでも 6 割が判定可能であるとの予測を得ることができた.

しかし、図 2 の latin1、utf8 の”なし”に該当する LSE は、本手法では判別が不可能という結果となった．解決策として、例えばカテゴリ化し、その中で新たに固有 BNP を作成、新たに得られた固有 BNP による判別方法も考えられるが、検討が必要である.

参考文献

[1] Canasai Kruengkrai, Virach Sornlertlamvanich, and Hitoshi Isahara. Language, script, and encoding identification with string kernel classifiers. *Proceedings of the First International Conference on Knowledge, Information and Creativity Support Systems*, 08 2006.

[2] Tommi Vatanen, Jaakko J. Väyrynen, and Sami Virpioja. Language identification of short text segments with n-gram models. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pp. 3423–3430. European Language Resources Association (ELRA), 2010.

[3] Chew Y. Choong, Yoshiki Mikami, C A. Marasinghe, and S Nandasara. Optimizing n-gram order of an n-gram based language identification algorithm for 68 written languages. Vol. 2, pp. 21–28, 12 2009.