J-011

固定長バイト列一次元スペクトルを利用した高速言語判別法

A method for fast language identification using one-dimensional spectrum of fixed length byte sequences

高野 凱 † 中平 勝子 † 北島 宗雄 †
Gai Takano Katsuko T. Nakahira Muneo Kitajima

1 はじめに

近年，デジタルデバイドが問題になっている．デジタルデバイドとは情報格差のことであり，”デジタルデバイドは物理的な通信手段の有無のみをもって測ることはできず，デジタル技術を扱う能力やデジタル技術によってもたらされる効用といった段階においても格差が生じている”[1] となっている．また言語間においてはより一層の格差が広がっている現状がある．現存する言語は 6000 以上あると言われているが，その中でインターネット空間上で表現することが出来る言語は一部となっている．つまり言語の表現の問題でインターネット空間上の情報技術の恩恵を取得できない人々が存在する．そのため, テーマとして言語間のデジタルデバイドを取り扱う．扱うにあたって言語間の格差の現状を知る必要がある．現状を知るためにはインターネット上にある様々な言語で書かれたテキストを集計する必要がある．しかし，集計されたテキストがどの言語で書かれているかが明らかではないことがあり，ここに言語判定の必要性が示された．よって本稿では言語間格差の調査の前段階として，言語判定エンジンの作成を目指す．

Chew 等による先行研究 [2] によって言語判定エンジンの実装は既に行なわれている．しかし，現在の言語判定エンジンでは解析対象と教師データを全て比較している．そのため今後インターネットがさらに発達し解析対象のデータ量が増加した場合，言語判定エンジンとして実用的な時間で解析を行えない可能性があり，今後を見据えるとより高速で処理を行えるよう改善する必要性がある．

2 LSE による言語判別

ウェブ上には複数の文字，複数の文字コードを持つ言語が存在する．言語判別に向けて，これらを区別して取り扱う必要が有るため，Language-Script-Encode の頭文字を取った LSE を判別することを本稿の言語判別とする．この LSE を判別するための手法を以下に説明する．まず，LSE 毎の辞書を用いる方式がある．あらかじめ言語ごとの辞書を作成しておき，判別対象の文字・単語を辞書と照らし合わせて判別を行う手法である．高精度な判定が期待できるが，LSE 毎に大量のデータが登録されている辞書の必要性があり，今回扱う多言語の場合にはそもそも辞書の作成自体が困難である事が予想される．

次にテキストの N-gram を取る方式がある．教師データとして，単語を区別せずテキスト中から指定した数の文字列の頻度を求める．判別対象も同様に頻度を求め，これの相関を取ることで言語判定が可能となる．しかし本稿では 500 以上もの LSE を扱うため，それぞれのテキストの文字コードやフォント等を用意する必要がある．また 1 文字が何バイトで表わされるかは LSE によって異なっているため，切り出した文字列が持つ情報量が不定である．そして上記を応用した 16 進化テキストの N-gram がある．テキストの時と同様に行うが，テキストを 16 進数化してから N-gram を取得する．この際，1 文字を 1 バイトと置き換え，N-gram を行う．文字列で言語判定の処理を行おうとする場合，読み込む際にそれぞれの文字コードを用意してテキストに合ったエンコード処理を行う必要が生じる．しかし．バイト列で処理を行うことによって，文字コードによらず 16 進数という統一されたフォーマットで言語判定処理を一括に行うことが可能になる．対象テキストを 16 進数化するのは非常に容易であり，この手法は LSE によらず，判定を行えるといった利点がある．よって本稿では 16 進化テキストを用いて言語判別を行う事とする．

N-gram 方式には 0 頻度問題といったものが存在しており，教師データにデータが無いからといってその組み合わせが存在しない訳ではないといった様な事があり得る．この点に十分注意して N-gram を使用する必要がある．また，言語判定の際に用いる教師データは，同一の内容で３００以上の多言語で翻訳されているかつ，言語判定を行うにあたって適当なテキストサイズであるとして，「世界人権宣言」を用いる．

3 固定長バイト列一次元スペクトル

固定長バイト列の一次元スペクトルを分光法を用いて定義する．

分光法とは，物理的観測量の強度を周波数，エネルギー，時間などの関数として示すことで，対象物の定性・定量あるいは物性を調べる科学的手法である．例として，ある試料に光を照射し，透過光や反射光の強度を測定する．この強度は物質固有のパターンであるため，これにより試料の物質同定が行える．

本稿の言語判定エンジンでは，16 進化テキストの N-gram から抽出された固定長バイト列の出現頻度を正規化した数値を強度，固定長バイト列を物理的観測量と捉えることによって．分光法と同様の方法で言語を判定する．このように定義した結果，固定長バイト列スペクトルがどのように出現するかは図 1 の様に予想できる．1 は頻度の分布がある一箇所が集まる様子を表している．分布が集まるという事は，同じバイト列の出現が多いということになる．つまり言語全体の文字数が少ない場合，そのような分布になると考えられる．2 は頻度の分布がばらついている様子を表している．分布がばらつくのは多様な文字でテキストが表現されているためであり，このような分布に

† 長岡技術科学大学

なる言語は言語全体の文字数が多い言語であると考えられる．3 は分布がばらついているが頻度が非常に多いバイト列を持っている様子を表している．そのような言語はある特定の文字を多用する傾向があると考えられる．例として，英語，日本語，ラオ語の教師データのスペクトルを図 2 に示す．橙色は英語，緑色が日本語，紫色がラオ語となっており，それぞれ橙が 1，緑が 2，紫が 3 のパターンを表している．

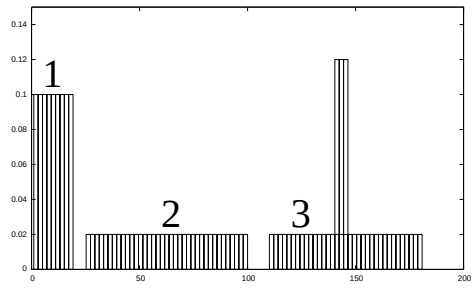

図 1　バイト列出現予想図

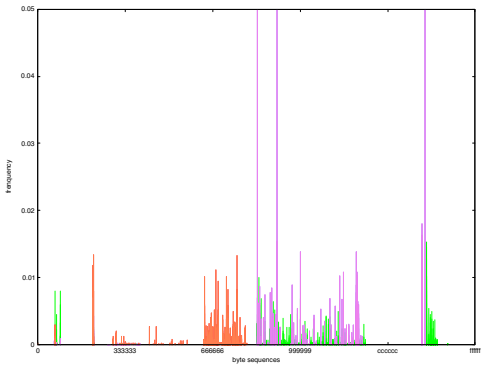

図 2　教師データのスペクトル図

分布の様子を確認するために，教師データのテキストを 16 進化し，N-gram を用いて解析を行った．その結果，言語によってはその言語にしか現れない固有なバイト列の存在を確認することができた．固有なバイト列が存在する言語数について以下の表 1 に示す．本稿では教師データの性質によるノイズも考慮し，固有なバイト列が 10 以上ある教師データは判別性能を持っている事とする．表 1 から，bi に比べて tri が固有バイト列を多く持っている事が確認できる．また，固有バイト列を 10 種類以上持つ教師データの数についても同様のことが言え，その数はおよそ 3 倍ほどになった．このことから，tri を用いて判定を行うと 376 の LSE について一様に判定を行える事が示された．よって固定長バイトとして 3 バイトを採択する．

表 1　教師データが持つ固有なバイト列

N-gram	固有バイト列数	判別性能を持つ教師データ数
bi	8762	118/572
tri	99995	376/572

言語固有なバイト列の存在と有用性が確認されたので，これを用いた言語判別の概略図を，図 3 に示す．固有バイト列を 10 以上持つ LSE については，一意に言語判定が行えると考えられるため，固定バイト列用のマッチングデータを作成し，判別に用いる．固有バイト列が 10 より少ない，またそもそも固有なバイト列が存在しない LSE については，あらかじめ作成する言語種別-固定長バイト列の教師データ行列 $\boldsymbol{T}$ と，判別対象のテキストから得た固定長バイト列ベクトル $\boldsymbol{s}$ の乗算によって，結果を示すことが出来る．$\boldsymbol{T}$ が m 行 n 列で表される行列とし，教師行列の出現頻度を f_{mn}，判別対象の出現頻度を f_{s_n} とすると，ある LSE である確率 p_n は以下の式 1 で表される．教師データとして用いた全ての LSE を行方向に置き，tri において表現出来る全ての固定長バイト列を行として置く．先行研究において作られていた，言語判別エンジンとは違い，固定長バイト列が存在する 376 の LSE では単純なマッチングで判別でき，固定長バイト列が存在しない場合には，行列の計算によって判別が行える，さらにこの行列演算において，対象テキストに無いバイト列については計算を省略する事が出来るため，高速な言語判別が行えると考えられる．

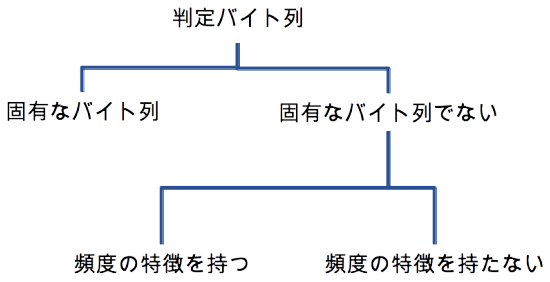

図 3　言語判定概要図

$$\boldsymbol{T}\cdot\boldsymbol{s}=\begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{mn} \end{pmatrix}\begin{pmatrix} f_{s_1} \\ f_{s_2} \\ \vdots \\ f_{s_n} \end{pmatrix}=\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix} \quad (1)$$

4　まとめと今後の課題

言語判定手法を比較した結果，N-gram を 16 進数を用いて拡張した手法を用いて言語判定を行う．分光法を用いることで，従来の言語判定エンジンより計算量が削減できることを示した．今後は言語判定エンジンの実装と，性能判定を行う必要がある．

参考文献

[1] 上島 智大, 中平 勝子，三上 喜貴：デジタルデバイドの評価指標について一提案

[2] Yew Choong Chew, Yoshiki Mikami, Robin Lee Nagano：Language Identification of Web Pages Based on Improved N-gram Algorithm, International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, 47p

[3] 三上 喜貴, 中平 勝子，児玉茂昭：言語天文台からみた世界の情報格差