4N-11

Shift Codon Matching を用いた言語判別に関する一考察

和田祐一† 岩橋政宏† 中鉢欣秀†† 三上喜貴†††
†長岡技術科学大学 電気系 ††科学技術振興機構
†††長岡技術科学大学 経営情報系
e-mail : (ywada@tech, iwahashi@vos, chubachi@oberon, mikami@kjs).nagaokaut.ac.jp

1. はじめに

世界には 6000 以上もの言語が存在するにも係らず，ネットワーク上で不自由なく使用されているものは，そのごく一部であり，中には文字コードが未開発の言語も存在する．現在，この「言語間デジタルデバイド」の解消を目的として，Web 上のテキストを自動的に収集･判別して言語別活動量を推定し，蓄積した関連基礎データを内外に提供するといった研究が行われている[1]．この研究における言語の判別方法としては Shift Codon Matching という手法が利用されているが，他に考えられる判別手法との比較はこれまであまり報告されていない．そこで本稿では，Shift Codon Matching の多値処理への拡張，Bayes 決定則，確率分布の類似度尺度といった各手法との比較検討を行い，得られた知見を報告する．

2. Shift Codon Matching を用いた言語判別

Shift Codon Matchingとは符号化された文字コードのバイト列をベースとするN-Gram分布を用いた判別手法である[1]．この手法では判別対象文章のシフトコドンu_j（j=1,2,…,k）から，(1)式により評価値を算出する．そして言語iでの評価値が 0.95 以上，且つその他の言語での評価値が 0.92 未満の時のみ言語iを判別結果とする手法である．

$$T_i=\sum_{j=1}^{k}s_i(u_j)/k,\quad s_i(u_j)=\begin{cases}1 & (u_j\in S_i)\\ 0 & (Otherwise)\end{cases} \tag{1}$$

但し，S_iは言語iの教師データである．また，本稿では文献[1]と同様にN=3 とした．

3. 各種判別法

3.1 多値処理への拡張

Shift Codon Matching とは, (1)式からも分かるように, 判別する Shift Codon が教師データに含まれている(“1”), いない(“0”)の二値処理であると考えることができる．ここで一般的には，二値処理に比べて自由度の高い多値処理の方が判別精度は上がると考えられるため，本稿では次のような処理を行い，多値処理へと拡張を試みる．

$$T_i=\sum_{j=1}^{k}s_i(u_j)/k,\quad s_i(u_j)=\begin{cases}M_j & (u_j\in S_i)\\ 0 & (Otherwise)\end{cases} \tag{2}$$

$$M_j=ceil\left(freq_j/\max\_freq_i\times step\right)$$

但し，$freq_j$はShift Codon u_jの出現頻度，max_freq_iは言語iのShift Codonの最大出現頻度，$step$は最大階調値，$ceil$(・)は正の無限大方向への丸めを表す．

3.2 Bayes 決定則

この方法は期待損失の最小化により評価を行う．ここで, 期待損失$L(\omega_j\mid x)$は次式により定義される．

$$\begin{aligned}L(\omega_j\mid x)&=\sum_{i=1}^{c}l(\omega_j\mid\omega_i)P(\omega_i\mid x)\\&=1-P(\omega_j\mid x)\end{aligned} \tag{3}$$

$$l(\omega_j\mid\omega_i)=\begin{cases}0 & (j=i)\\ 1 & (j\neq i)\end{cases}$$

但し，$P(\omega_i\mid x)$はパターンxが観測された時，クラスω_iに属する確率，cはクラス数である．従って，(3)式より，期待損失の最小化は事後確率$P(\omega_j\mid x)$の最大化と等価であることが分かる．ここで,各クラス間の生起確率を等確率と仮定すると, Bayesの定理より, 結局は$P(x\mid\omega_j)$の大小関係により判別が行われることになる．

3.3 確率分布の類似度尺度

確率分布関数間の類似度尺度としては，Tankard により次の関数が提案されている[2].

$$Tankard(P,Q)=\sum_{x}\left|P(x)-Q(x)\right| \tag{4}$$

但し，$P(x)$, $Q(x)$は N-Gram 分布の確率分布関数である．この関数は確率分布が完全に一致する場合 0 となり，全く一致しない場合 2 となる．従って，この評価値が最も小さいものとして判別が行われることになる．

4. 評価実験

4.1 実験仕様

判別手法の評価尺度としては，線形判別法のクラス内分散・クラス間分散比最大基準を用いる．教師

A study on language distinction with Shift Codon Matching.
†Y.WADA, M.IWAHASHI, Nagaoka Univ. of Technology, Department of Electrical Engineering
††Y. CHUBACHI, Japan Science and Technology Agency
†††Y.MIKAMI, Nagaoka Univ. of Technology, Department of Management and Information System Engineering

データとしては世界人権宣言(UDHR)から復元抽出した 10000, 5000, 1000 個の Codon を用いる．また，学習データとしては UDHR から別途復元抽出した 1000，100 個の Codon を用いる．評価は，各条件下で独立に 1000 回の試行を繰り返し, 得られた判別結果に対して，クラス内分散・クラス間分散比を求めて比較を行った．クラス内分散・クラス間分散比は大きくなる程に，より判別し易い結果であると言えるため，クラス内分散・クラス間分散比が大きい判別法程，判別性能が高い手法であると考えられる．

4.2 多値化処理による評価結果

学習量及び解析量が各々10000，1000 の条件下における多値化処理による評価結果を Fig.1 に示す．Fig.1 において, 最大階調値=1 が Shift Codon Matching に相当する．評価結果としては，一部で Shift Codon Matching を上回る結果が得られたものの，全体としては階調数が増加する程に，クラス内分散・クラス間分散が低下し, 判別性能が低下する傾向となった．但し，ここで言う階調数とは最大階調値に階調値が零の場合を加えたものであるため，最大階調値に 1 を足した数となる．

4.3 Bayes 決定則との比較結果

Bayes 決定則を用いて判別を行った結果を Table .1, Table .2 に示す．結果は英語-スペイン語と英語-インドネシア語で傾向が異なり，英語-スペイン語では Shift Codon Matching の方が良い結果を示しているのに対し，英語-インドネシア語では，Bayes 決定則の方が良い結果を示した．この理由としては，Bayes 決定則は学習した Codon の確率分布に依存した判別法であるため，言語間の関係が近い組み合わせ程効果が得られないのではないかと予想される．

4.4 Tankard を用いた判別法との比較結果

Tankard を用いて判別を行った結果を Table .1, Table .2 に示す．結果は解析 Codon 数により傾向が異なり，Codon 数が多い場合には，Tankard の方が良い結果となり，逆に少ない場合には Shift Codon Matching の方が良い結果となった．この理由としては，Codon 数が少ない場合には有意な確率分布が得られないためであると考えられる．

5. まとめと今後の課題

本稿では Shift Codon Matching を用いた言語判別の多値処理への拡張及び他手法との比較・検討を行った．多値化処理の結果としては，言語の組み合わせに応じて最適な階調数が存在し，中には Shift Codon Matching を上回る組み合わせも存在することが明らかとなった．しかしながら，全体としては階調数が増加する程に，判別性能が低下する結果となった．また，他手法との比較では，言語間の距離が近いものに限っては Bayes 決定則を用いることで，より良い判別が可能であることを明らかにした．この手法の適用可能性を探るためにも，各種言語間のグルーピングが今後の課題として挙げられる. 更に，解析 Codon 数が十分に与えられる場合には，確率分布の類似度尺度を用いることで，他の手法を上回る解析が可能であるということを明らかにした．

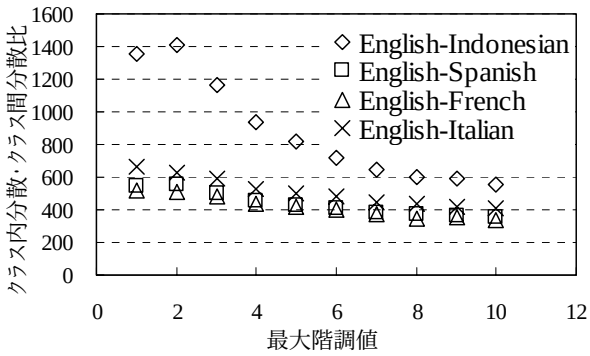

Fig.1 多値化処理による評価結果

Table.1 評価結果(English Latin1-Spanish Latin1)

学習量	10000		5000		1000	
解析量	1000	100	1000	100	1000	100
Shift Codon Matching	542	55	516	53	349	34
Bayes 決定則	437	44	396	40	257	25
Tankard	899	50	819	48	401	36

Table.2 評価結果(English Latin1-Indonesian Latin1)

学習量	10000		5000		1000	
解析量	1000	100	1000	100	1000	100
Shift Codon Matching	1351	128	1287	129	690	70
Bayes 決定則	1718	181	1536	163	738	77
Tankard	1735	81	1632	82	864	63

参考文献

[1] I.SUZUKI, et.al., “A Language and Character Set Determination Method Based on N-gram Statistics,” ACM Trans. on Asian Language Information., vol.1, no.3, pp269-278, Sept. 2002.

[2] 松浦司, 金田康正, “近代日本小説家 8 人による文章の n-gram 分布を用いた著者判別,” 自然言語処理研究会 137-1, June 2000.