1 日本語テキストを n-gram で学習する

日本語テキストを対象として、3-gram を作ってみます。
「はじめての機械学習に掲載されているプラグラムは、シフト JIS の環境、主にウィンドウズ環境では、期待通りの結果が出なかったので、「はじめての AI プログラミング」小高知宏著からのプログラムを引用しました。

2 日本語テキストの 3-gram を作成するプログラム make3gram.c

```
/*
  make3gram.c
  sjis の全角文字で記述されたファイルについて、
  3gram を作成するプログラム
  入力行は標準入力から与え、出力は標準出力に出力します。
  使い方
  ./make3gram <（入力ファイル名）>（出力ファイル名）
  入力ファイルには、テキストファイルを指定します。
  出力ファイルには、3gram を出力します。
*/

#include <stdio.h>
#define MAX 65535 * 3

int getsource(char *s)
{
  int n = 0;

  while((s[n++] = getchar()) != EOF);

  return n;
}

void getwidechar(char *t, char *s, int n)
{
  int in   = 0;
  int out = 0;
  int d;

  while(in < n){
     d = (unsigned char)s[in];
     if(((d > 0x7F) && (d < 0xA0)) || (d > 0xDF) && (d < 0xF0)){
       t[out++] = s[in++];
       t[out++] = s[in++];
     }else{
       ++in;
```

```
    }
  }
  t[out] = '\0';
}

void outputtarget(char *target)
{
  int i = 0;
  while((target[i] != '\0') && (target[i+2] != '\0') && (target[i+4] != '\0')){
    putchar(target[i++]);
    putchar(target[i++]);
    putchar(target[i]);
    putchar(target[i+1]);
    putchar(target[i+2]);
    putchar(target[i+3]);
   putchar('\n');
  }
}

int main(int argc, char *argv[])
{
  char source[MAX];
  char target[MAX];
  int    numchar;

  numchar = getsource(source);

  getwidechar(target, source, numchar);

  outputtarget(target);

  return 0;
}
```

3 「make3gram」の操作方法

「make3gram」の操作方法の例

```
./make3gram < ningen.txt > ningen35gram.txt
```

上記の「./」は、Linux でのプログラムを実行する時の例です。
ウィンドウズでは、make3gram < ningen.txt > ningen3gram.txt「Enter」
と入力してください。

4 「make3gram」の実行結果の例

「make3gram」の実行結果の例
人間失
間失格
失格太
格太宰
太宰治
宰治【
治【テ
【テキ
テキス
キスト