階層 N-gram マッチ

天野晃

平成 22 年 12 月 13 日

1 N-gram とは

N-gram とは、文字列を長さ N のサブ文字列に分解したときの、そのサブ文字列である。たとえば、”N-gram”という文字列の、延べ 2-gram は、{”N-”,”-g”,”gr”,”ra”,”am”} である。複数の文字列へ拡張可能で、これを次の式で表す (ベクトル:サブ文字列)。

$$G(n, st1, st2, ...) \tag{1}$$

さらに、その総出現数を次の式で表す (スカラ:0 もしくは自然数)。

$$F(n, st1, st2, ...) \tag{2}$$

さらに、各要素における出現数を次の式で表す (スカラ:0 もしくは自然数)。

$$F(e_i|n, st1, st2, ...) \tag{3}$$

また、このとき、以下を満たしているべきである。

$$F(n, st1, st2, ...) = \sum_i F(e_i|n, st1, st2, ...) \tag{4}$$

パラメータ等　パラメータとしては、オーバーラップの長さ (前例では 1)、端部サブ文字列にワイルドカードを追加するか (前例では追加していない)、などがある。

2 N-gram マッチとは

N-gram マッチとは、ふたつ以上の文字列 ($st1$、$st2$、$st3$、...) より生成されるそれぞれの N-gram において全てにマッチ (出現) する要素があることを言い (一般的に n-gram co-occurrence と呼ぶものであると思われる)、その数を次の式で表す (スカラ:0 もしくは自然数)。

$$M(n, st1, st2, st3, ...) \tag{5}$$

	a	a	x
a	1	1	0
b	0	0	0
x	0	0	1

図 1: 1-gram マッチの例

”a”と”a”のマッチは 2 回、”x”と”x”のマッチは 1 回、全てのマッチの合計は 3 回。

さらに、各要素におけるそれを次の式で表す (スカラ:0 もしくは自然数)。

$$M(e_i|n, st1, st2, st3, ...) \tag{6}$$

また、このとき、以下を満たしているべきである。

$$M(n, st1, st2, st3, ...) = \sum_i M(e_i|n, st1, st2, st3, ...) \tag{7}$$

および

$$M(e_i|n, st1, st2, st3, ...) = F(e_i|n, st1)\ F(e_i|n, st2)\ F(e_i|n, st3)\ ... \tag{8}$$

たとえば、$st1$ =”aax”、$st2$ =”abx”というふたつの文字列の 1-gram マッチの数 ($M(n, st1, st2)$) は、3($st1$ のふたつの”a”が $st2$ の”a”とマッチ、双方の”x”がマッチ、計 3 回マッチ) となる (図 1)。

3 (N-gram を基にした) 類似度の定義

N-gram を基にした複数の文字列 ($st1$、$st2$、$st3$、...) 間の類似度を S とする。

3.1 定義.A

S を次のように定義する (スカラ:実数)。

$$S(n, st1, st2, st3, ...) = \frac{M(n, st1, st2, st3, ...)^2}{M(n, st1, st1)\ M(n, st2, st2)\ M(n, st3, st3)\ ...} \tag{9}$$

たとえば、前述の $st1$、$st2$ 間の類似度は図 2 のようになる。

問題点　少なくともひとつの種類の、直観に反するような例が見付かっている: $M(1,$”abx”,”dex”) と $M(1,$”aax”,”bbx”) においては、いずれも文字列長が同じで、前後者ともマッチする文字はひとつのみであるが、類似度が異なる (前者は 0.11、後者は 0.04 : 図 3)。

$$S(1, st1, st2) \quad = \quad \frac{\begin{pmatrix} & \mathrm{a} & \mathrm{a} & \mathrm{x} \\ \mathrm{a} & 1 & 1 & 0 \\ \mathrm{b} & 0 & 0 & 0 \\ \mathrm{x} & 0 & 0 & 1 \end{pmatrix}^2}{\begin{array}{cccc} & \mathrm{a} & \mathrm{a} & \mathrm{x} \\ \mathrm{a} & 1 & 1 & 0 \\ \mathrm{a} & 1 & 1 & 0 \\ \mathrm{x} & 0 & 0 & 1 \end{array} \times \begin{array}{cccc} & \mathrm{a} & \mathrm{b} & \mathrm{x} \\ \mathrm{a} & 1 & 0 & 0 \\ \mathrm{b} & 0 & 1 & 0 \\ \mathrm{x} & 0 & 0 & 1 \end{array}} \quad = \quad \frac{9}{15}$$

図 2: 定義.A による N-gram ベースの類似度の例

自乗と掛け算は文字 (gram) がマッチした回数に対して行う。分子を自乗することにより値はつねに 0–1 となる。

$$\frac{\begin{pmatrix} & \mathrm{a} & \mathrm{b} & \mathrm{x} \\ \mathrm{d} & 0 & 0 & 0 \\ \mathrm{e} & 0 & 0 & 0 \\ \mathrm{x} & 0 & 0 & 1 \end{pmatrix}^2}{\begin{array}{cccc} & \mathrm{a} & \mathrm{b} & \mathrm{x} \\ \mathrm{a} & 1 & 0 & 0 \\ \mathrm{b} & 0 & 1 & 0 \\ \mathrm{x} & 0 & 0 & 1 \end{array} \times \begin{array}{cccc} & \mathrm{d} & \mathrm{e} & \mathrm{x} \\ \mathrm{d} & 1 & 0 & 0 \\ \mathrm{e} & 0 & 1 & 0 \\ \mathrm{x} & 0 & 0 & 1 \end{array}} \quad \neq \quad \frac{\begin{pmatrix} & \mathrm{a} & \mathrm{a} & \mathrm{x} \\ \mathrm{b} & 0 & 0 & 0 \\ \mathrm{b} & 0 & 0 & 0 \\ \mathrm{x} & 0 & 0 & 1 \end{pmatrix}^2}{\begin{array}{cccc} & \mathrm{a} & \mathrm{a} & \mathrm{x} \\ \mathrm{a} & 1 & 1 & 0 \\ \mathrm{a} & 1 & 1 & 0 \\ \mathrm{x} & 0 & 0 & 1 \end{array} \times \begin{array}{cccc} & \mathrm{b} & \mathrm{b} & \mathrm{x} \\ \mathrm{b} & 1 & 1 & 0 \\ \mathrm{b} & 1 & 1 & 0 \\ \mathrm{x} & 0 & 0 & 1 \end{array}}$$

図 3: 定義.A による直観とは異なる N-gram ベースの類似度の例

どちらもマッチする文字は x のみであるが、分母のマッチ数が異なるため異なる類似度を返す。

$$S(1, st1, st2) \quad = \quad 6 \times \frac{\begin{array}{cccc} & \text{a} & \text{a} & \text{x} \\ \text{a} & 1/3 & 1/3 & 0 \\ \text{b} & 0 & 0 & 0 \\ \text{x} & 0 & 0 & 1/2 \end{array}}{3 \quad \times \quad 3} \quad = \quad \frac{7}{9}$$

図 4: 定義.B による N-gram ベースの類似度の例

6 は”aax”と”abx”の、3 は、”aax”または”abx”の文字の総出現数。a-a、x-x のマッチに対し、$\frac{1}{3}$、$\frac{1}{2}$ と、重み付けが行われている。

3.2 定義.B

S を次のように定義する (スカラ:実数)。

$$S(n, st1, st2, ...) = F(n, st1, st2, ...) \ \frac{\sum_i \left(M(e_i|n, st1, st2, ...) \times F(e_i|n, st1, st2, ...)^{-1} \right)}{F(n, st1) \ F(n, st2) \ ...} \qquad (10)$$

たとえば、前述の $st1$、$st2$ の類似度は図 4 のようになる。

問題点　後に述べるように階層化が容易ではない。

4 階層 N-gram マッチ

階層 N-gram マッチとは、以上のような単語レベルのマッチングあるいは類似度の計算を、さらに上位のフレーズ、センテンスレベルへ階層的に積み上げることを言う。

4.1 定義.A による階層 N-gram マッチ

定義.A による階層 N-gram マッチでは、ワード間の N-gram マッチを行った後、フレーズ間の”N-word”マッチを行い、その後に、センテンス間の”N-phrase”マッチを行う... というように階層的にマッチングを行なう。また、階層を越える度に、マッチの値に対し閾値を設け 0 または 1 にする (必須ではない)。たとえば、$s1$ =”aax abx abc axx”、$s2$ =”aax abx abx axx”というフレーズ、1-gram/2-word/閾値 0.5 の場合、図 5 のように行う。

4.2 定義.B による階層 N-gram マッチ

定義.B によるマッチングの階層化はスマートには行えず、複雑な手順、もしくは、場当たり的なものとならざるを得ない。

$$
\begin{pmatrix} & \text{aax} & \text{abx} & \text{abc} & \text{axx} \\ \text{aax} & 1 & 0.6 & 0.27 & 0.64 \\ \text{abx} & 0.6 & 1 & 0.44 & 0.6 \\ \text{abx} & 0.6 & 1 & 0.44 & 0.6 \\ \text{axx} & 0.64 & 0.6 & 0.67 & 1 \end{pmatrix}^2
$$

$$
\begin{array}{c|cccc} & \text{aax} & \text{abx} & \text{abc} & \text{axx} \\ \text{aax} & 1 & 0.6 & 0.27 & 0.64 \\ \text{abx} & 0.6 & 1 & 0.44 & 0.6 \\ \text{abc} & 0.27 & 0.44 & 1 & 0.67 \\ \text{axx} & 0.64 & 0.6 & 0.67 & 1 \end{array} \times \begin{array}{cccc} \text{aax} & \text{abx} & \text{abx} & \text{axx} \\ 1 & 0.6 & 0.6 & 1 \\ 0.6 & 1 & 1 & 0.6 \\ 0.6 & 1 & 1 & 0.6 \\ 1 & 0.6 & 0.6 & 1 \end{array} \rightarrow
$$

$$
\begin{pmatrix} & \text{aax} & \text{abx} & \text{abc} & \text{axx} \\ \text{aax} & 1 & 1 & 0 & 1 \\ \text{abx} & 1 & 1 & 0 & 1 \\ \text{abx} & 1 & 1 & 0 & 1 \\ \text{axx} & 1 & 1 & 1 & 1 \end{pmatrix}^2
$$

$$
\begin{array}{c|cccc} & \text{aax} & \text{abx} & \text{abc} & \text{axx} \\ \text{aax} & 1 & 1 & 0 & 1 \\ \text{abx} & 1 & 1 & 0 & 1 \\ \text{abc} & 0 & 0 & 1 & 1 \\ \text{axx} & 1 & 1 & 1 & 1 \end{array} \times \begin{array}{cccc} \text{aax} & \text{abx} & \text{abx} & \text{axx} \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{array} \rightarrow \frac{4^2}{3 \times 9} = \frac{16}{27}
$$

図 5: 定義.A による 2 階層 N-gram 類似度の例

ワードのマッチに 1-gram/閾値 0.5、フレーズのマッチに 2-word、を用いた。

$$\underbrace{\begin{pmatrix} & \text{aax} & \text{abx} & \text{abc} & \text{axx} \\ \text{aax} & 1 & 0.78 & 0.44 & 0.89 \\ \text{abx} & 0.78 & 1 & 0.67 & 0.78 \\ \text{abx} & 0.78 & 1 & 0.67 & 0.78 \\ \text{axx} & 0.89 & 0.78 & 0.33 & 1 \end{pmatrix}}_{3\times 3} \rightarrow \underbrace{\begin{pmatrix} & \text{aax} & \text{abx} & \text{abc} & \text{axx} \\ \text{aax} & 1 & 1 & 0 & 1 \\ \text{abx} & 1 & 1 & 1 & 1 \\ \text{abx} & 1 & 1 & 1 & 1 \\ \text{axx} & 1 & 1 & 0 & 1 \end{pmatrix}}_{3\times 3} = \frac{7}{9}$$

図 6: 定義.B-2 による 2 階層 N-gram 類似度の例

ワードのマッチに 1-gram、フレーズのマッチに 2-word/閾値 0.5、を用いた。

定義.B-1: (前掲の例で、) まず、aax-abx、abx-abc、abc-axx、... と 3×3 の行列内のセルに 2×2 の行列を含む形式の表を作成する。次に各セルの値を、対角要素を足して 2 で割ったものとする。こうして出来た表に定義.B を最適応する。

定義.B-2: 第二階層以上では、閾値以上で連続して n 回マッチするその回数を総計し、探索空間の積で割る。定義.B-2 によるマッチの例を図 6 に示す。