d-bigram 情報を用いた統語的規則の抽出

慶應義塾大学大学院理工学研究科計算機科学専攻
佐藤 健吾 堤 純也 中西 正和

1 はじめに

近年、統計情報を用いた自然言語処理が盛んになっているが、単語の文脈中での使われ方あるいは分布の情報を用いて自動的に単語をクラスタリングする手法は科学的、実用的な見地から注目を集めている。このような手法は、語彙の獲得、統計的な言語モデルの構築、統計的手法における希薄なデータなどの問題に対して有効であると思われる。現在ではこのような手法を応用して、統計情報から統語規則を獲得する研究が行われている。本稿では、近年考案された d-bigram [1, 2] を用いて、単語のクラスタリングを試み、その結果に基づいた統語規則の抽出について考察する。

2 d-bigram

2.1 d-bigram モデル

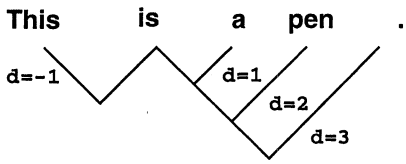

図 1: d-bigram の例

図 1 のように 2 単語が $d \in \mathbf{Z}$ だけ離れて出現する確率モデルのことを **d-bigram** と呼ぶ。Ω をコーパス中の単語の集合、$c(v,w,d)$ を単語 v,w が d だけ離れて出現した回数 $(v,w \in \Omega)$ とすると、単語 x,y の d に関する d-bigram $P(x,y,d)$ は以下のように定義される [1, 2]。

$$P(x,y,d) = \frac{c(x,y,d)}{\sum_{v,w\in\Omega} c(v,w,d)} \qquad (1)$$

この値から、単語 x から d だけ離れて単語 y が現れることが他の単語と比べてどれだけもっともらしいかということがわかる。

2.2 d-bigram による単語間の相互情報量の定義

相互情報量は 2 つのイベントの相互依存量として考えられたものである。ある 2 つの事象 x,y の相互情報量 $I(x,y)$ は一般に以下のように定義される。

$$I(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

ここで $p(x)$ は事象 x が起こる確率、$p(x,y)$ は事象 x と y が同時に起こる確率である。

d-bigram による単語間の距離 d に関する相互情報量は以下のように定義される。

$$MI_d(x,y) = \log \frac{P(x,y,d)}{P(x)p(y)} \qquad (2)$$

これにより単語 x から d だけ離れて単語 y が出現することがどれだけ情報を持っているかを測ることが可能となる [1, 2, 4]。

2.3 d-bigram による単語間の類似度の定義

直感的に考えて、単語間の類似度の性質としては同じような使われ方をする単語同士の類似度がなるべく高くなるべきである。このような単語を発見するために、d-bigram のデータを用いた単語の特徴ベクトルを定義する [1, 2]。

$$\begin{aligned} v(w) = & \\ & (c(w,w_0,d_0),,\ldots,c(w,w_{N-1},d_0), \\ & c(w,w_0,d_1),\ldots,c(w,w_{N-1},d_1), \\ & \vdots \\ & c(w,w_0,d_{m-1}),\ldots,c(w,w_{N-1},d_{m-1})) \end{aligned} \qquad (3)$$

ただし $d_i \in \mathbf{D} = \{\ldots, -2, -1, 1, 2, \ldots\}$,

$$m = |\mathbf{D}|,\ N = |\Omega|$$

このベクトルは単語 w を中心にしたすべての $d \in \mathbf{D}$ に対するコーパス中のすべての単語の出現頻度を並べたもので、このベクトルが似ていれば「同じような使われ方」といえるだろう。

このことを数量的に表すために、2 つの単語の特徴ベクトルの角度を計算し、これを単語間の類似度とする [1, 2]。

$$D_v(w_1, w_2) = \arccos \frac{v(w_1) \cdot v(w_2)}{|v(w_1)||v(w_2)|} \quad (4)$$

3 統計情報による統語規則の獲得

統計情報から統語規則を獲得する方法は大きく分けて 2 種類の研究が現在行なわれている。一つは、文脈自由文法のような品詞列の生成規則をコーパスから抽出する方法、もう一つは、格文法のようにある動詞の目的語にくる名詞のクラスをコーパスから抽出する方法である。本稿では、前者のアプローチで d-bigram 情報による統語規則の獲得を試みた。

3.1 n-gram 統計情報による品詞列の生成規則の獲得

森 [3] は、品詞タグ付きコーパスから n-gram を用いて品詞列の生成規則を獲得する手法を提案している。この手法は次の 2 つを仮定している。

- 生成規則の右辺に現れる品詞列は環境から独立して高い出現頻度で現れる。
- 同じ非終端記号から直接導出される品詞列は類似した環境を持つ。

アルゴリズムは次の通りである。

1. n-gram と (n+1)-gram から計算された左右のエントロピー H_l, H_r を用いて独立な品詞列を抽出する。

$$H_l(\boldsymbol{pos}) = -\sum_i Pl_i(\boldsymbol{pos}) \log Pl_i(\boldsymbol{pos})$$
$$H_r(\boldsymbol{pos}) = -\sum_i Pr_i(\boldsymbol{pos}) \log Pr_i(\boldsymbol{pos}) \quad (5)$$

ここで $\boldsymbol{pos}$ は品詞列、$Pl_i(\boldsymbol{pos})$ は $\boldsymbol{pos}$ の左側に品詞 pos_i が現れる確率、$Pr_i(\boldsymbol{pos})$ は $\boldsymbol{pos}$ の右側に品詞 pos_i が現れる確率である。$H_l + H_r$ が大きい品詞列は独立な品詞列であると考える。

2. 品詞列の左右に現れる品詞の確率分布ベクトルの距離を類似度として用い 1. で抽出した品詞列を以下の手順に従いクラスタリングする。
   (a) 独立な品詞列のすべての組合せについて類似度を計算する。
   (b) ノードに品詞列、アークに品詞列間の類似度が対応するグラフを生成する。
   (c) ある閾値以下の値のアークを消去する。
   (d) グラフの連結成分で品詞列を分類する。
3. 生成されたクラスタに新しい非終端記号を割り当て、各々のクラスタの構成要素からクラスタに対応する非終端記号への生成規則を作る。
4. コーパス中の独立な品詞列を 3. で割り当てた非終端記号に置き換える。
5. 1. に戻る。

以上のことを新しい独立な品詞列が抽出できなくなるまで続ける。

森はこの手法により、人間が手作業で作った文法と比べて妥当であるような生成規則を獲得することが可能であるとしている。

4 d-bigram 情報を用いた統語的規則の抽出

本稿では、統計量として 2 節で述べた d-bigram を採用し、3.1 節で述べたアルゴリズムと同等のことを行なうことによって統語的規則を抽出することを試みた。

4.1 独立な品詞列の抽出

4.1.1 相互情報量による文の切り分け

まずはじめに、延澤の手法 [4] を用いて文を繋がりやすい品詞列に分解する。この手法は入力文

$pos_1, pos_2, \ldots$ について品詞 pos_i と pos_{i+1} 間の有繋評価値 $UK(i)$ を式 (2) を用いて計算し、閾値を下回る i で文を切り分けるものである。

$$UK(i) = \sum_{d=1}^{d_{max}} \sum_{j=i-(d-1)}^{i} \frac{MI_d(pos_j, pos_{j+d})}{g(d)} \quad (6)$$

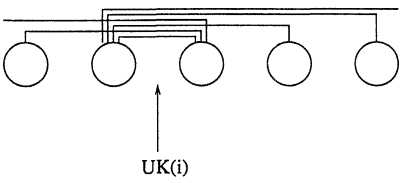

図 2: 有繋評価値の計算

4.1.2 エントロピーによる独立な品詞列の抽出

延澤の手法で得られた品詞列は、3.1 節で述べた独立な品詞列の仮定を満たしているとは限らないため、さらに式 (5) を求めることで独立な品詞列を抽出する。

d-bigram モデルでは、$\boldsymbol{pos} = pos_j pos_{j+1} \ldots$ と pos_i が連続して現れる回数は $\boldsymbol{pos}$ の長さが 1 でない限り得ることができないため、$pos_k \in \boldsymbol{pos}$ と pos_i の d-bigram から近似計算して $Pl_i(\boldsymbol{pos})$ と $Pr_i(\boldsymbol{pos})$ を求める。

4.2 品詞列のクラスタリングによる生成規則の抽出

得られた独立な品詞列に対して 3.1 節で述べた方法と同様にしてクラスタリングを行なう。ただし、品詞列の類似度として式 (4) を用いる。

d-bigram モデルでは $\boldsymbol{pos}$ の長さが 1 でない限り $c(\boldsymbol{pos}, pos_i, d)$ を得ることができないため、$c(pos_k, pos_j, d)$ (ただし $pos_k \in \boldsymbol{pos}$) を用いて近似計算する。

このようにして得られたクラスタに属する品詞列は同じ非終端記号に書き換えられると考えられるので、ある非終端記号を導入し、クラスタに属する品詞列からその非終端記号への生成規則が存在するとみなす。

4.3 生成規則による品詞列の置換

ここまでで、ある終端記号から非終端記号への生成規則を抽出することが可能となったが、実際の文法には非終端記号からの生成規則も存在する。このような生成規則を抽出するために、ここまでに得られた生成規則でコーパスを書き換え、その結果に対してここまでの手順を繰り返す。

5 実験結果および考察

4 節で述べた方法を用いて生成規則の抽出を試みた。実験に用いたコーパスは SUSANNE Corpus [5] である。このコーパスは約 15 万語からなり、各単語に品詞タグが付いている。本稿の実験ではこの品詞タグを利用した。

実験の結果、途中経過を含めて以下のような出力を得ることができる。

- 独立な品詞列
- 生成規則
- コーパスの置換結果

5.1 独立な品詞列

図 3 に独立な品詞列を抽出した結果を示す。

```
ii at jj nn
np
at nn cc
nn iw at jj nn
```

図 3: 抽出された独立な品詞列の一部

結果を見る限り、独立であると思われる品詞列が抽出されているが、独立な品詞列がすべて抽出されているとは限らないため、有繋評価値、左右のエントロピーの閾値の調節を行ない最適化する必要があると思われる。

5.2 生成規則の獲得

図 4 に獲得した生成規則の一部を示す。

```
sym01-0 --> ii at jj nn
sym01-0 --> ii at jj nnj
sym01-0 --> ii at jj nnl
sym01-0 --> nn ii at nnl
sym01-0 --> nn io at nnl
```

図 4: 生成規則の一部

生成規則の妥当性に関する評価方法は一般的には確立されていないため客観的な評価は行なっていないが、妥当であると思われる結果もいくつか含まれている。生成規則の獲得はクラスタリングの結果が大きな影響を与えるため、閾値の最適化が必要であると思われる。

5.3 コーパスの置換

図 5 にコーパスのある 1 文の置換結果の一部を示す。

```
at np nnl jj nn vvd npd at nn io
np gg jj jj nn vvd
yil at nn yir cst dd
nn vvd nnl yf
```

```
at sym01-4 sym01-23 vvd npd at nn io
sym01-44 gg jj sym01-23 vvd
sym01-49 at nn yir sym01-75 sym01-60
nn vvd sym01-33 yf
```

図 5: コーパスの置換結果の一部

今回の実験で用いた置換のアルゴリズムは、単純に文の先頭から適用できる規則で置換していくものであるが、得られた生成規則を最適に適用しているものはあまり多くなかった。コーパスの置換は次の生成規則抽出に非常に大きな影響を与えるので、置換アルゴリズムの改善が必要であると思われる。

5.4 全体の考察

この手法は大きく分けて 3 段階のフェーズからなり、それぞれに高い精度が要求される。今回の実験では、クラスタリング、コーパスの置換の精度があまりよくなかったため、全体としての結果はあまりよいものとは言えない。

6 今後の課題

今後の課題としては以下のようなことがあげられる。

- 定性的な評価
- 森の手法との比較
- 閾値の最適化
- 近似計算の改善
- より正確なクラスタリングアルゴリズムの適用
- より精度の高い置換アルゴリズムの適用

参考文献

[1] 堤 純也, 新田 朋晃, 小野 孝太郎, 延澤 志保. 統計情報を用いた多言語間機械翻訳システム. 人工知能学会研究会, pages 7–12, 1993.

[2] Tsutsumi, J., Nitta, T., Ono, K., Nobesawa, S. and Nakanishi, M. Multi-lingual Machine Translation Based on Statistical Information. *Qualico*, pages 147–152, 1994.

[3] 森 信介. 統計によるコーパスからの統語規則の獲得. 修士論文, 京都大学大学院工学研究科, 1995.

[4] 延澤 志保. 自然言語における有繋文字列の抽出. 言語処理学会第 2 回年次大会, 3 1996.

[5] University of Sussex, School of Cognitive & Computing Sciences. *The SUSANNE Corpus*, 11 1994. ftp:sable.ox.ac.uk:/pub/ota/public/susanne.