d-bigram を用いた単語のクラスタリング

1H－4

佐藤 健吾, 堤 純也, 孫 大江, 延澤 志保, 佐野 智久, 中西 正和
慶應義塾大学理工学研究科計算機科学専攻

1. はじめに

単語の使われ方による単語の自動的な分類の手法は、科学的、実践的見地から興味が持たれている。例えば、

- 言語学的な構造の分布や語彙の獲得における疑問に対する、精神的あるいは計算的な学習の展望にどのような関係があるか。
- どうやって希薄なデータをうまく処理したり、統計的な言語モデルを生成するか。

といった問題が挙げられる。

巨大なコーパスになると、大部分が sparse データになってしまうため、信頼性が低くなってしまうことが良く知られている。この問題に対して sparse データを“似ている”イベントで代用することが考えられるが、類似度を単語のクラスや対応するモデルの生成に直接使う方法は明らかになっていない [1]。

本論文では、d-bigram [2] を用いた単語のクラスタリングの方法について考察し、どのようなクラスタが生成されるかを検証する。

2. d-bigram

2.1 d-bigram モデル

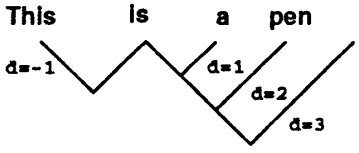

図 1: d-bigram の例

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 [2]。

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

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 [3]。

$$\begin{aligned} v(w) = & \\ & (c(w,w_0,d_0),,\ldots,c(w,w_{N-1},d_0), \\ & c(w,w_0,d_1),\ldots,c(w,w_{N-1},d_1), \\ & \qquad\vdots \\ & c(w,w_0,d_{m-1}),\ldots,c(w,w_{N-1},d_{m-1})) \end{aligned} \tag{2}$$

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 [3]。

$$D_v(w_1,w_2) = \arccos \frac{v(w_1)\cdot v(w_2)}{|v(w_1)||v(w_2)|} \tag{3}$$

Word Clustering Using D-bigram
Kengo SATO, Junya TSUTSUMI, Da Jiang SUN, Shiho NOBESAWA, Tomohisa SANO, Masakazu NAKANISHI
Department of Computer Science, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa Pref., 223, Japan

3. クラスタリングアルゴリズム

2.2 で述べたように単語間の類似度を定義すると、多変量解析の技法の一つであるクラスタリングにより単語の分類を行うことができる。

現在までに、クラスタリングのためのアルゴリズムが数多く提案されてきたが、その一つに図 2 のようなデンドログラムを生成する階層的手法がある。この方法では、すべての対象がそれだけで大きさ 1 のグループをなしている状態から出発する。そして近いグループをしだいに併合してゆき、ついにはすべての個体が 1 つにグループになるという過程を経る。

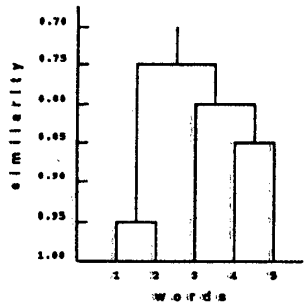

図 2: デンドログラムの例

4. 実験方法

対象とするコーパスは、Brown Corpus と呼ばれているもので、総単語数約 1,200,000 語、語彙数約 45,000 語、約 57,000 文の英語の例文からなる。

今回の実験ではクラスタリングの対象とする単語数を出現頻度上位およそ 1,000 語とした。これらの単語で Brown Corpus のおよそ 74 % をカバーすることができる。

また、式 (2) の特徴ベクトル $v(w)$ の定義において、各要素 $c(w, w_n, d_m)$ の w_n は、w を特徴付ける単語であるが、出現頻度が多いほど w の特徴を多く表すと考えられる。そこで、w の特徴を良く表しているであろう出現頻度上位の単語のみを用いて特徴ベクトルを構成し、単語間の類似度の計算の効率化を図る。ここでは、Brown Corpus の半分以上 (60%) をカバーする出現頻度上位 200 語のみを特徴ベクトルの構成に用いた。

5. 結果

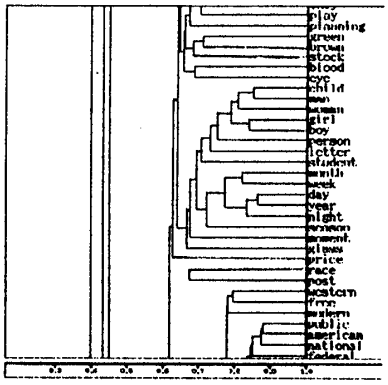

図 3: クラスタリング結果の一部 (1)

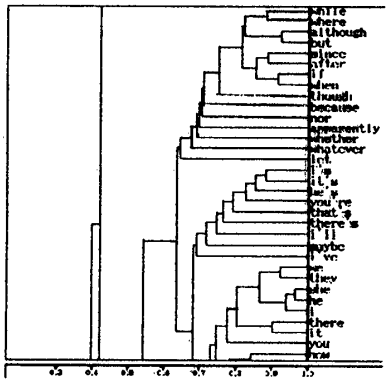

図 4: クラスタリング結果の一部 (2)

生コーパスの情報のみから「人間の生活に深く関係のある名詞」「数字を表す名詞」「人間の記憶に関係のある動詞」など意味的な概念 (図 3) や、一般にいわれている「接続詞」「代名詞」「前置詞」といった文法的な情報 (図 4) を抽出することができた。また、動詞の現在形、過去形の違いを抽出することもできた。

参考文献

[1] Pereira, F., Tishby N. and Lee L. Distributional clustering of English words. *Proceedings of ACL-93*, June, 1993.

[2] 堤 純也, 新田 朋晃, 小野 孝太郎, and 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.

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