単語の半教師ありクラスタリング

鍛治伸裕　　喜連川優

東京大学 生産技術研究所
{kaji, kitsure}@tkl.iis.u-tokyo.ac.jp

1　はじめに

単語クラスタリングは，自然言語処理や情報検索の多くの場面で利用される基礎技術である．これまでに，様々な単語クラスタリング手法が提案されているが，いずれも教師なし学習法が用いられている．これに対して本論文では，半教師あり学習法 [2] を用いて単語クラスタリングを行うことを提案する．さらに，語彙統語パターンを用いてコーパスから類義語を獲得して，教師データを自動生成する方法もあわせて提案する．

2　教師なし単語クラスタリング

まず，教師なし学習法を用いた代表的な手法として，分布類似度 [3] にもとづく単語 (=名詞) クラスタリングを考える．これは，係り受け関係にある動詞を用いて名詞の類似度を定義する方法である．係り受け関係にある動詞との共起頻度を用いると，名詞 n は以下の特徴ベクトル $\boldsymbol{\phi}(n)$ で表すことができる．

$$\boldsymbol{\phi}(n) = (f_{nv_1}, f_{nv_2}, \cdots f_{nv_V}) \tag{1}$$

ただし，f_{nv} は名詞 n と動詞 v の共起頻度で，V はコーパスにおける動詞の異なり数である．分布類似度にもとづく単語クラスタリングとは，すなわち $\boldsymbol{\phi}(n)$ と $\boldsymbol{\phi}(n')$ が類似していれば n と n' を同一クラスタに割り当てるという方法である．

教師なし単語クラスタリングは混合分布モデルを用いて定式化できる．$\boldsymbol{\phi}(n)$ が混合多項分布から生成されたと仮定すると，名詞 n の確率は

$$p(n) = \sum_{z=1}^{Z} p(\boldsymbol{\phi}(n)|z)p(z) = \sum_{z=1}^{Z} \frac{(\sum_v f_{nv})!}{\prod_v f_{nv}!} (\prod_v \mu_{vz}^{f_{nv}})\pi_z$$

と定義される．μ_{vz} は多項分布のパラメータ，π_z は混合比である．すなわち $\sum_z \mu_{vz} = 1$，$\sum_z \pi_z = 1$ を満たす．同様に，名詞集合 $\mathbf{n} = \{n_i\}_{i=1}^N$ の確率は，$\mathbf{n}$ に対応する隠れ変数を $\mathbf{z} = \{z_i\}_{i=1}^N$ とすると

$$p(\mathbf{n}) \quad = \quad \sum_{\mathbf{z}} p(\mathbf{n}|\mathbf{z})p(\mathbf{z}) \tag{2}$$

と定義できる．ただし $p(\mathbf{n}|\mathbf{z}) = \prod_{i=1}^N p(n_i|z_i)$，$p(\mathbf{z}) = \prod_{i=1}^N p(z_i)$ である．モデルのパラメータは EM アルゴリズムを用いて推定することができる．

3　半教師あり単語クラスタリング

次に，教師データを取り込めるように，前節で説明したモデルを拡張する．ここでは，教師データとして類義語対が与えられたという状況を想定する．名詞 n_i と n_j が類義であるということは，言い換えると，隠れ変数 z_i と z_j が同じ値をとるということである．すなわち，今考えている教師データとは，隠れ変数間の制約と等価である．そこで以下の議論では，教師データは隠れ変数間の制約 $\mathbb{C}$ の形で与えられる仮定とする．各制約 $\langle i, j\rangle \in \mathbb{C}$ は，z_i と z_j が同じ値をとることを意味する．さらに制約違反に対するペナルティ $w_{ij}(>0)$ が対応づけられているものとする．

教師データを取り込むため，Basu らにならって，混合分布モデルの $p(\mathbf{z})$ を以下のように変更したモデルを用いる [2]．

$$p(\mathbf{z}) \quad = \quad \prod_{i=1}^{N} p(z_i) \times \frac{1}{G} \exp\{-\sum_{\langle i,j\rangle \in \mathbb{C}} \delta(z_i \neq z_j) w_{ij}\}$$

$\delta(\cdot)$ はデルタ関数，G は正規化項である．制約が破られるとデルタ関数が 1 となることから，制約を破るような値が出現しにくくなっていることが分かる．

モデルのパラメータは，通常の混合分布モデルと同様に EM アルゴリズムで推定する．しかし，隠れ変数同士が独立でないため，E-step の計算が困難である．そこで，$p(\mathbf{z})$ の近似分布を平均場近似で求め，その近似分布を使って E-step の計算を行う．平均場近似で

は $p(\mathbf{z})$ を $q(\mathbf{z}) = \prod_{i=1}^{N} q_i(z_i)$ で近似する．$q(\mathbf{z})$ のパラメータは真の分布 $p(\mathbf{z})$ との KL divergence が最小となるものを選ぶ．パラメータ $q_{ik} = q_i(z_i = k)$ は以下の更新式で計算できる [4].

$$q_{ik}^{(t+1)} \propto p(n_i, k) \exp\{-\sum_{j \in \mathcal{N}_i} (1 - q_{jk}^{(t)}) w_{ij}\} \qquad (3)$$

ただし $\mathcal{N}_i = \{j | \langle i, j \rangle \in \mathbb{C}\}$ であり，$q_{ik}^{(t)}$ は t 回目の繰り返しにおける q_{ik} の値である．

4 制約の導出

語彙統語パターンを用いて類義語対をコーパスから自動獲得し，そこから隠れ変数間の制約を導出する．実験では以下の 4 種類の語彙統語パターンを用いた．

X や Y　X も Y も　X と Y と　X, Y,

しかし，単純にパターンにマッチした語を収集したのでは適合率に問題があるため後処理を行った．後処理の基本的な考え方は次のようなものである．まず，パターンで収集された類義語対はグラフとみなすことができる (語が頂点，類義関係が辺)．そして，密な辺を持つ頂点集合 (典型的にはクリーク) は信頼できる類義語集合であると考えられる．そこで，このグラフから連結度の高い頂点集合を抽出して使うことにした．

連結度の高い頂点集合は次のような手続きで求めた．基本的な処理はボトムアップクラスタリングと同じである．まず最初に，各語が大きさ 1 のクラスタを形成しているような状態を作る．そして，適当な順番で 2 つのクラスタを選び，2 つを併合して新たに得られるクラスタの連結度が高ければ，その 2 つを併合する．ここでは，全ての頂点が残りの頂点の過半数と辺で結ばれているクラスタを，連結度が高いクラスタとした．これにより，連結度の高い頂点集合が 1 つのクラスタにまとめられる．この処理を，併合できるクラスタが存在しなくなるまで繰り返した後，大きさが S 以上のクラスタ[1]を信頼できる類義語集合として取り出す．最後に，同じ類義語集合に属する名詞の間に制約を与える．重みは全て w とする．

5 実験

ウェブと新聞記事から収集した約 2 億文を用いて実験を行った．クラスタ数は 1,000 とした．クラスタリング結果は，日本語語彙大系 [5] をもとにテストデータを作成して，B-CUBED アルゴリズム [1] で求めた適合率と再現率で評価した．制約の重み w は 100 とした．これは，テストデータと同様の方法で作成したディベロップメントデータを使って最適な値を求めた結果である．ディベロップメントデータとテストデータは，ともに 400 名詞 (20 カテゴリ) から成る．表 5 に実験の結果を示す．教師データを用いることによって，単語クラスタリングの精度が向上していることが確認できる．

	適合率	再現率	F 値
教師なし	76.8	35.3	48.4
半教師あり	79.3	36.3	49.8

表 1: 実験結果

6 まとめ

本論文では，単語クラスタリングに半教師あり学習法を用いることを提案した．また，語彙統語パターンを用いて教師データを自動生成する手法もあわせて提案した．実験の結果，教師データを用いることによってクラスタリングの精度が向上することを確認した．

参考文献

[1] Amit Bagga and Breck Baldwin. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of ACL*, 1998.

[2] Sugato Basu, Mikhail Bilenko, and Raymond J. Mooney. A probabilistic framework for semi-supervised clustering. In *Proceedings of SIGKDD*, pp. 59–68, 2004.

[3] Donald Hindle. Noun classification from predicate-argument structure. In *Proceedings of ACL*, pp. 268–275, 1990.

[4] Tilman Lange, Martin H.C. Law, Anil K. Jain, and Joachim M. Buhmann. Learning with constrained and unlabelled data. In *Proceedings of CVPR*, pp. 731–738, 2005.

[5] 白井諭 横尾昭男 中岩浩巳 小倉健太郎 大山芳史 林良彦池原悟（編）. 日本語語彙大系. 岩波書店, 1997.

[1] 実験では $S = 5$ とした．