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

半教師あり学習法の説明をする前に，まずは教師なし学習法を用いた単語クラスタリングについて述べる．ここでは代表的な手法として，分布類似度 [5] 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．

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

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

$$p(n) = \sum_{z=1}^{Z} p(\boldsymbol{\phi}(n)|z) \times p(z) \tag{2}$$

$$= \sum_{z=1}^{Z} \frac{(\sum_v f_{nv})!}{\prod_v f_{nv}!} (\prod_v \mu_{vz}^{f_{nv}}) \times \pi_z \tag{3}$$

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．このモデルで隠れ変数 z_i は名詞 n_i の意味カテゴリであると解釈することができる．

同様に，名詞集合 $\mathbf{n} = \{n_i\}_{i=1}^N$ の確率も定義できる．$\mathbf{n}$ に対応する隠れ変数を $\mathbf{z} = \{z_i\}_{i=1}^N$ とし，これらが互いに独立であると仮定すると

$$p(\mathbf{n}) = \sum_{\mathbf{z}} p(\mathbf{n}|\mathbf{z}) p(\mathbf{z}) \tag{4}$$

となる．ただし，$p(\mathbf{n}|\mathbf{z}) = \prod_{i=1}^N p(\boldsymbol{\phi}(n_i)|z_i)$，$p(\mathbf{z}) = \prod_{i=1}^N p(z_i)$ 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．E-step と M-step の計算は以下の通りである[1]．

E-step

$$p(z_i = k|n_i) = \frac{(\prod_v \mu_{vk}^{f_{n_i v}})\pi_k}{\sum_z (\prod_v \mu_{vz}^{f_{n_i v}})\pi_z} \tag{5}$$

M-step

$$\mu_{\gamma k} = \frac{\alpha + \sum_i f_{n_i \gamma} p(z_i = k|n_i)}{\alpha V + \sum_v \sum_i f_{n_i v} p(z_i = k|n_i)} \tag{6}$$

$$\pi_k = \frac{\alpha + \sum_i p(z_i = k|n_i)}{\alpha Z + \sum_z \sum_i p(z_i = z|n_i)} \tag{7}$$

ここで重要なのは E-step の計算である．E-step が上記のように効率的に計算可能なのは，隠れ変数が互いに独立だからであるという点に注意されたい．

3　半教師あり単語クラスタリング

前節のモデルに教師データを取り込むことを考える．ここでは，教師データとして類義語対が与えられたと

[1] M-step の α はスムージング係数である．実験では α=1.0 とした．

いう状況を想定する．名詞 n_i と n_j が類義語対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みなすことができる．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し，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．

半教師あり単語クラスタリングの確率モデルは，前述のモデルとほぼ同じである．唯一の違いは $p(\mathbf{z})$ を以下のように変更する点である [2].

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

$\delta(\cdot)$ 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 となるため，$p(\mathbf{z})$ をこのように変更することによって，制約を破るような値が出現しにくくなる．

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．M-step の計算は前節のものと全く同じであるが，E-step の計算が異なる．半教師あり学習では隠れ変数同士が独立でないため，E-step で以下の計算を行わなくてはならない．

$$p(z_i|\mathbf{n}) = \sum_{\mathbf{z}_{\neg i}} p(\mathbf{z}_{\neg i}, z_i|\mathbf{n}) \quad (8)$$

$$\propto \sum_{\mathbf{z}_{\neg i}} p(\mathbf{n}|\mathbf{z}_{\neg i}, z_i) p(\mathbf{z}_{\neg i}, z_i) \quad (9)$$

ここで $\mathbf{z}_{\neg i}$ は z_i 以外の全隠れ変数の集合を指す．$p(z_i|\mathbf{n})$ を直接求めることは，計算量の問題から難しい．そこで，Lange らと同じく平均場近似を用いて E-step の計算を行う [6]．平均場近似では $p(\mathbf{z}|\mathbf{n})$ を $q(\mathbf{z}) = \prod_{i=1}^{N} q_i(z_i)$ で近似する．この近似分布を用いると，E-step の計算は

$$p(z_i|\mathbf{n}) \simeq \sum_{\mathbf{z}_{\neg i}} q(\mathbf{z}_{\neg i}, z_i) = q_i(z_i) \quad (10)$$

となる．近似分布 $q(\mathbf{z})$ のパラメータは真の分布 $p(\mathbf{z}|\mathbf{n})$ 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，以下の更新式にもとづく反復法で求めることができる [6].

$$q_{ik}^{(t+1)} \propto p(n_i, k) \exp\{-\sum_{j \in \mathcal{N}_i} (1 - q_{jk}^{(t)}) w_{ij}\} \quad (11)$$

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．

一般に半教師あり学習では，制約なしデータと制約ありデータの数の偏りが，モデルの性能に悪影響を与えることがある．この問題は，制約ありデータの尤度に重み付けを行うことで回避できる [7, 6]．制約なし / ありデータの対数尤度をそれぞれ L_u と L_c をおくと，これまでに説明した EM アルゴリズムは $L_u + L_c$ の局所最適解を求めていることに相当する．これに対して，重み付けを行った場合では $L_u + \lambda L_c$ を最適化する．$\lambda(>0)$ は，制約ありデータの重要度を示すハイパーパラメータである．重み付けされた対数尤度は，EM アルゴリズムの M-step を以下のように修正したアルゴリズムで最適化することができる．

$$\mu_{\gamma k} = \frac{\alpha + \sum_i \Lambda_i f_{\gamma n_i} p(z_i = k|n_i)}{\alpha V + \sum_v \sum_i \Lambda_i f_{v n_i} p(z_i = k|n_i)}$$

$$\pi_k = \frac{\alpha + \sum_i \Lambda_i p(z_i = k|n_i)}{\alpha Z + \sum_z \sum_i \Lambda_i p(z_i = z|n_i)}$$

ただし，制約なしデータの場合は $\Lambda_i = 1$ で，制約ありデータの場合は $\Lambda_i = \lambda$ である．

4 制約の導出

通常，教師データは人手で作成される．しかし，本論文では，単語クラスタリングのための教師データを自動生成する方法を提案する．すでに議論したように，制約 $\langle i,j\rangle \in \mathbb{C}$ が存在するということは，名詞 n_i と n_j が類義語であることと等価である．そこで，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，それをもとに制約を自動導出する．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 (図 1)．そこで，このグラフから連結度の高い頂点集合だけ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

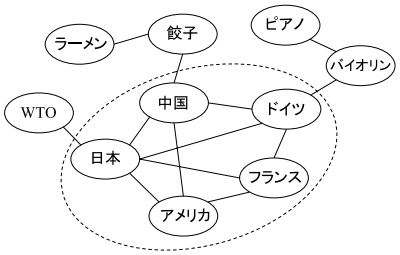

図 1: 単語グラフの例．連結度の高い頂点集合が破線で囲まれている．

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[2]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．表 1 に実験で獲得した類義語集合の例を示す．

表 1: 類義語集合の例

カボチャ，キュウリ，ナスビ，カブ，ニンジン
風邪，ぜんそく，花粉症，皮膚炎，湿疹
冷蔵庫，洗濯機，掃除機，乾燥機
ソニー，日立，シャープ，松下，三菱，東芝
東京，京都，大阪，名古屋，奈良，千葉，埼玉

5 実験

データ 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．全ての文を Juman と KNP を用いて解析して，名詞と動詞の係り受けを抽出した[3]．低頻度語を削除したのち，8,700 万の係り受けを得た．係り受け，名詞，動詞の異なりは，それぞれ 6,048,779，100,273，36,910 である．

4 節で述べた語彙統語パターンをこの 2 億文に適用したところ，97,296 の類義語対を獲得することができた．そして後処理を施した結果，364 の類義語集合 (名詞の異なり数は 2,499) を得た．実験では，これをもとに生成した制約を使った．

[2]実験では $S = 5$ とした．

[3]http://nlp.kuee.kyoto-u.ac.jp/nl-resource/

テストデータは既存のシソーラスをもとに作成した．日本語語彙大系 [9] から 20 カテゴリを無作為に選び，各カテゴリから 20 語ずつ (合計 400 語) 取り出し，これをテストデータとした．同様にして，別の 20 カテゴリからディベロップメントデータも作成した．

実験結果 クラスタリングの結果は，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．表 2 に，教師なし学習を用いた場合と，半教師あり学習を用いた場合の結果を示す．EM アルゴリズムの初期値依存性を考慮して，実験結果は 5 回の試行の平均値をのせている．この結果から，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．なお，ハイパーパラメータ w と λ は，$w \in \{1, 10, 100\}$ と $\lambda \in \{1, 1.2, 1.4, 1.6, 1.8, 2.0\}$ の中から，ディベロップメントデータを使って最適な値を求めた．また，クラスタ数は 1,000 に固定した．

表 2: 実験結果

	w	λ	適合率	再現率	F 値
教師なし	–	–	76.8	35.3	48.4
半教師あり	100	1.0	79.3	36.3	49.8

次にハイパーパラメータがクラスタリング精度に与える影響を調べた．図 2 は，テストデータにおける適合率と λ の関係をプロットしたものである．図中の semi-1，semi-10，semi-100 というのは，w が 1，10，100 のときの結果で，unsupervised が教師なし学習の結果である．同様に，図 3 に再現率と λ の関係を示す．これらの図から，少なくとも本実験で用いたデータについては，適切な w の選択が重要であることが分かる．一方，λ は w ほど深刻な影響を与えていない．w さえ適切に選べば，ほとんどの場合で適合率と再現率ともに教師なし学習法を上まわっている．

6 関連研究

Bollegala らは，SVM による教師あり学習法を用いて単語間の類似度を求めている [3]．Bollegala らの研究と比較したとき，本研究の特徴は，教師データがクラスタリングの目的関数に直接とりこまれている点である．すなわち，Bollegala らの手法は metric learning であり，我々の手法は constrained clustering であると言うことができる [2]．

半教師あり単語クラスタリングは set expansion と

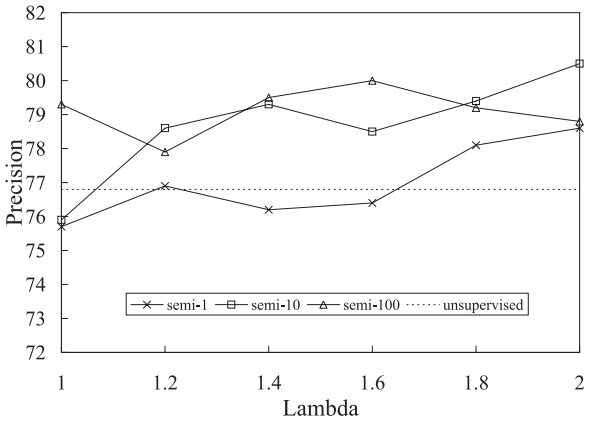

図 2: 適合率と λ

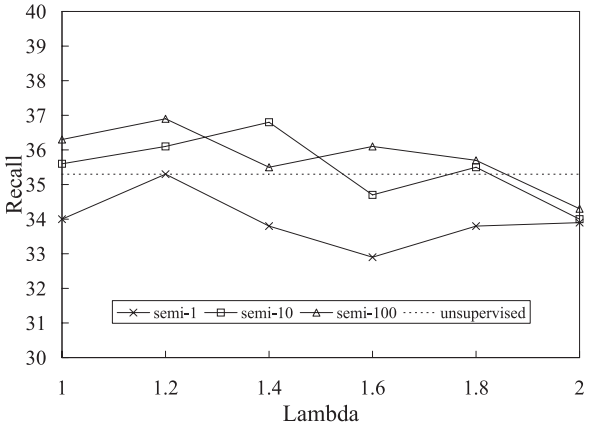

図 3: 再現率と λ

関連が深い [4, 8]．Set expansion とは，あるクラスタに所属する少数の単語集合がクエリとして与えられて，そのクラスタに属するその他の単語を検索するタスクである．これは我々の議論しているクラスタリングとは問題設定がやや異なるものの，同一クラスタのメンバーを探すために教師データ (i.e., クエリ) を利用しているという点で類似性がみられる．

7 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．今後は，応用システムの中で提案手法の有効性を検証していく予定である．また，能動学習の適用なども今後の課題としたい．

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] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. An integrated approach to measuring semantic similarity between words using information available on the web. In *Proceedings of NAACL*, pp. 340–347, 2007.

[4] Zoubin Ghahramani and Katherine A. Heller. Bayesian sets. In *Proceedings of NIPS*, pp. 14–21, 2005.

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

[6] 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.

[7] Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, Vol. 39, No. 2/3, pp. 103–134, 1999.

[8] Richard C. Wang and William W. Cohen. Language-independent set expansion of named entities using the web. In *Proceedings of ICDM*, pp. 1015–1021, 2007.

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