2008－NL－184 (2)
2008／3／27

単語の用例の半教師有りクラスタリング

杉山 一成† 奥村 学†

† 東京工業大学 精密工学研究所
〒 226–8503 神奈川県横浜市緑区長津田町 4259
E-mail: †sugiyama@lr.pi.titech.ac.jp, †oku@pi.titech.ac.jp

あらまし 単語の用例をクラスタリングすることは，教師有りの語義曖昧性解消手法のためのタグ付きコーパス作成支援，新語義候補の抽出，多義性解消システムの精度改善，などに有効であると考えられる．本研究では，この単語の用例のクラスタリングに，半教師有りクラスタリングを適用する．我々の提案する半教師有りクラスタリングは，種用例間に導入する制約に関して，“cannot-link” の制約を重視していること，また，語義タグを付与した種用例を含むクラスタの重心の変動を抑えること，において新規性がある．本論文では，この提案手法を「SENSEVAL-2 日本語辞書タスク」のデータに適用した結果について報告する．

キーワード 半教師有りクラスタリング，単語の用例，語義曖昧性解消

Semi-Supervised Clustering For Word Examples

Kazunari SUGIYAMA† and Manabu OKUMURA†

† Precision and Intelligence Laboratory, Tokyo Institute of Technology
4259 Nagatsuta, Midori-ku, Yokohama, Kanagawa, 226–8503 Japan
E-mail: †sugiyama@lr.pi.titech.ac.jp, †oku@pi.titech.ac.jp

Abstract Clustering for examples of a word is effective in supporting to construct tagged corpus for supervised word sense disambiguation, extracting candidates for a new word sense, improving accuracy in a word sense disambiguation system. In our study, we apply semi-supervised clustering approach to cluster examples of a word. Our proposed semi-supervised clustering approach is novel in that we focus on “cannot-link” with regard to constraints between seed examples and control the fluctuation of the centroid of a cluster. In this paper, we report the results obtained by applying our proposed method to the data of “SENSEVAL-2 Japanese dictionary task.”

Key words Semi-supervised clustering, Examples of a word, Word sense disambiguation

1. はじめに

単語の用例をクラスタリングすることは，図 1 に示すように，(a) 教師有りの語義曖昧性解消手法のためのタグ付きコーパス作成支援，(b) 新語義候補の抽出，(c) 語義曖昧性解消システムの精度改善，などに有効であると考えられる．なお，本論文では，図 1 中の「種用例」とは，語義タグを付与した単語の用例のことを表すものとする．

「(a) 教師有りの語義曖昧性解消手法のためのタグ付きコーパス作成支援」において，教師有りの語義曖昧性解消手法は，人手で語義のタグが付与された多数のデータを用いて，高い精度で語義の曖昧性解消を実現できる．しかしながら，この手法のために必要となる語義タグ付きコーパスを作成するためには，膨大な労力を必要とする．しかし，ある程度類似した用例をクラスタリングしてから語義タグを付与することができれば，このコストは格段に改善するものと考えられる．また，用例をクラスタリングする際には，クラスタ間に制約を与えることで，各クラスタには類似した用例が集まりやすくなることが期待される．「(b) 新語義候補の抽出」においては，クラスタリングを行なうことで，そのクラスタ内において，種用例から外れた部分に作成された要素は，新語義候補である可能性がある．また，「(c) 語義曖昧性解消システムの精度改善」においては，類似した用例が集められたクラスタ内において，素性を計算することで，語義を限定しやすくなり，語義の曖昧性解消に寄与できることが期待される．

しかし，単語の用例をクラスタリングする際，凝集型クラスタリングでは，クラスタリングを適切に導いていく基準がないために，正確なクラスタリングは難しい．また，これまでに提案されている半教師有りクラスタリングでは，制約を導入したり，距離を学習したりすることにのみ着目している．しかし，単語の用例に対して，半教師有りクラスタリングを適用し，精度の高いクラス

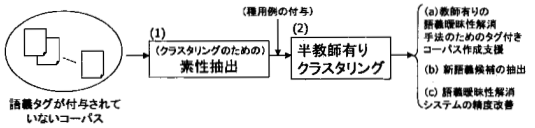

図 1 単語の用例に関する半教師ありクラスタリングの活用
Fig. 1 Usage of semi-supervised clustering for examples of a word.

タリング結果を得るためには，種用例を導入する方法，ならびに制約を導入する方法を工夫するとともに，種用例を含むクラスタの重心の変動を抑えることが重要であると考えられる．

本研究では，我々の提案する半教師有りクラスタリングを「SENSEVAL-2 日本語辞書タスク」のデータに適用した結果について報告する．

2. 関連研究

一般に，教師無しクラスタリングは，文書の組織化，閲覧，大規模文書の要約といったデータ解析を行なうために，重要な技術である．また，クラスタリングは，データへのラベル付けが実際的でなかったり，不可能であったりするような大規模なデータ集合を解析する処理において，有用である．このような教師無しクラスタリングは，何らかの教師情報を用いることによって，その精度が改善され得る．最近では，こうした何らかの教師情報を用いる，すなわち，半教師有りの手法でクラスタリングの精度を向上させることを目的とした研究が注目されている．

これまでに提案されている半教師有りクラスタリングの手法は，(1) 制約に基づいた手法，(2) 距離に基づいた手法，の二つに分類することができる．本章では，これらの研究について振り返る．

2.1 制約に基づいた手法

制約に基づいた手法は，ユーザが付与したラベルや制約を利用し，より適切にデータをクラスタリングできるようにする手法である．例えば，Wagstaff ら [1], [2] の半教師有り $K-$means アルゴリズムは，“must-link” (2 つの事例が同じクラスタに属さなければならない) と，“cannot-link” (2 つの事例が異なるクラスタに属さなければならない) という 2 種類の制約を導入し，これらの制約が侵されないことを保証して，データのクラスタリングを行なう．Basu ら [3] もまた，初期の種クラスタを生成し，クラスタリングを正確に行なうために，ラベル付きデータを利用する半教師有り $K-$means アルゴリズムを開発している．

2.2 距離に基づいた手法

距離に基づいた手法では，特定のクラスタリング尺度に基づいて，既存のクラスタリングアルゴリズムが利用される．しかし，このクラスタリング尺度を得るためには，教師付きデータにおけるラベルや制約を満たすための学習を必要とする．例えば，Klein ら [4] の研究では，類似した 2 点 (x_i, x_j) 間には “0”，類似していない 2 点間には $(\max_{i,j} D_{ij}) + 1$ と設定した隣接行列を作成して，クラスタリングを行なう．また，Xing [5] らの研究では，特徴空間の変換を行なうことで，マハラノビス距離を最適化する．さらに，Bar-Hillel ら [6] の研究では，適切な特徴には大きな重みを，そうでない特徴には小さな重みを与える RCA (Relevant Component Analysis) [7] により，特徴空間を変換する．

3. 提案手法

一般的な凝集型クラスタリングを，単語の用例のクラスタリングに適用した場合，クラスタリングを適切に導いていく基準がなく，正確なクラスタリングは難しい．一方，**2.** で述べた半教師有りクラスタリングは，制約を導入したり，距離を学習したりすることにのみ着目している．さらに，これらのアルゴリズムは，クラスタの重心の変動を抑えることを考慮していない．しかし，単語の用例に半教師有りクラスタリングを適用し，精度の高いクラスタリング結果を得るためには，種用例を導入する方法，ならびに制約を導入する方法を工夫するとともに，種用例を含むクラスタの重心の変動を抑えることが重要であると考えられる．この考えは，(1) “must-link” 導入時には，クラスタ内分散が大きくなり，正確なクラスタが生成されにくい，(2) 種用例を導入して半教師有りクラスタリングを行なう場合，通常の重心の計算法では重心の変動が大きくなり，クラスタリングの基準となる種用例を導入する効果が得られない，(3) 重心を完全に固定して半教師有りクラスタリングを行なう場合，その重心と類似度が高い用例しかマージされなくなり，多数の独立したクラスタが生成されやすくなる，という三つの観点に基づく．したがって，種用例を導入する方法，ならびに制約を導入する方法を工夫するとともに，種用例を含むクラスタの重心の変動を抑えることができれば，適切な用例が集められたクラスタが生成されるものと期待される．

以下，本章では，我々の提案する半教師有りクラスタリングについて説明する．本研究で提案する半教師有りクラスタリングの手法は，種用例間に制約を導入する方法，また，種用例を含むクラスタの重心の変動を抑えることにおいて，新規性がある．

3.1 クラスタリング時の素性

図 1(1) において，用例のクラスタリング時には，次の素性を導入した．

- 形態素素性
  - 対象単語および，前後 2 語までの単語 W の表記
    - $W(-2)$, $W(-1)$, $W(0)$, $W(+1)$, $W(+2)$
  - 対象単語および，前後 2 語までの単語の品詞 P，品詞細分類 P_d
    - $P(-2)$, $P(-1)$, $P(0)$, $P(+1)$, $P(+2)$
    - $P_d(-2)$, $P_d(-1)$, $P_d(0)$, $P_d(+1)$, $P_d(+2)$
- 構文素性
  - 対象単語が名詞の場合，その名詞が係る動詞
  - 対象単語が動詞の場合，その動詞のヲ格の格要素

括弧内の数値は，対象語の位置を “0” としたとき，その前 $(-2, -1)$，後 $(+1, +2)$ にある単語の位置を表す．なお，形態素解析器としては ChaSen(注1) を，構文解析器としては CaboCha(注2) を用いた．

本研究では，対象語 w に対するある用例 x の素性ベクトル $\boldsymbol{f}^x$ を式 (1) のように表す．

$$\boldsymbol{f}^x = (f_1^x, f_2^x, \cdots, f_n^x) \quad (1)$$

3.2 種用例，ならびに制約の導入法

単語の用例を高い精度でクラスタリングするためには，初期の種用例をどのように導入するかが重要になると考えられる．そこで本研究では，次の 3 つの手法で，半教師有りクラスタリングのための種用例を導入する．

(注1)：http://sourceforge.net/projects/masayu-a/
(注2)：http://sourceforge.net/projects/cabocha/

なお，以下において，対象語の用例のうち，種用例を選択するための用例を「訓練用例」と呼ぶこととする．**4.** で述べる実験では，この訓練用例の割合を変化させることで，提案手法のクラスタリング精度を検証した．

[手法 I] 訓練用例をクラスタリングして得られる各クラスタの重心を種用例とする．この際のクラスタリング手法は，K-means 法 [9] を用いた．

[手法 II] 訓練用例すべてを種用例とする．

[手法 III] 訓練用例に "KKZ" [10] と呼ばれる手法を適用し，種用例を生成する．

手法 I については，あるコーパスが与えられた場合に，数多くの単語の用例が存在する．したがって，あらかじめ，訓練用例をクラスタリングして代表となる点を見つけ，これらが異なる語義であると仮定し，"cannot-link" の制約を導入すれば，適切な用例が集まりやすくなることが期待される．手法 II については，すべてを種用例とした場合には，クラスタをマージしていく過程で，類似度の近い用例から，徐々に適切な用例を集めたクラスタが生成されるものと考えられる．手法 III は，あらかじめ，互いに距離の遠い用例を種用例として設定することで，クラスタどうしがマージされる際に，不適切なクラスタがマージされるのを防ぐ効果が期待される．なお，手法 III は，クラスタの初期化法について比較した研究 [11] において，高い精度のクラスタリング結果が得られる初期化法であると報告されている．この "KKZ" のアルゴリズムを図 2 に示す．

また，これらの 3 つの種用例の導入法において，次のような制約を導入する．上述したように，手法 I においては，クラスタの重心を代表点として選択しているので，これらが異なる語義であると仮定して，"cannot-link" の制約を導入する．また，手法 II，III においては，次の 4 種類の制約を導入する．

[制約 (a)] 異なる語義どうしに "cannot-link" を導入

[制約 (b)] 同じ語義どうしに "must-link" を導入

[制約 (c)] 異なる語義どうしに "cannot-link" を，同じ語義どうしに "must-link" を同時に導入

[制約 (d)] はじめに，異なる語義どうしに "cannot-link" を導入する．クラスタが生成されてきた段階 (今回はテスト用例の 3 割を超えた段階とした) で，クラスタどうしの重心間の距離が，設定した閾値 Th_{dis} よりも小さくなった場合に，"must-link" を導入する．ここで，$Th_{dis} = 0.597$ と設定した．

これらの制約は，次の観点に基づいて導入した．まず，制約 (a) については，クラスタ内に異なる語義の用例が含まれることを防ぐ効果が期待される．また，制約 (b) は，クラスタ内には同じ語義どうしの用例を集めやすくなることが期待される．制約 (c) においては，制約 (a)，制約 (b) の両方を考慮することで，クラスタ内に異なる語義の用例が含まれることを防ぎつつ，同じ語義どうしの用例が集まりやすくなることが期待される．また，制約 (d) においては，"cannot-link" の制約によって，クラスタ内に異なる語義の用例が含まれることを防ぎつつ，クラスタの状態が確定してきてから "must-link" の制約を導入することで，適切な用例が集められたクラスタどうしをマージすることができ，より正確なクラスタが生成されることが期待される．

さらに制約 (d) については，文献 [12], [13] に，半教師有りクラスタリングにおいては，"must-link" よりも "cannot-link" の制約を重視することが有効であること，また，文献 [14] において，制約を組合せ過ぎても精度の高いクラスタリング結果が得られない，と報告されていることにも基づく．図 3 に，この "must-link" 導入時のアルゴリズムを示す．

Algorithm: KKZ
Input: Set of features of examples $\boldsymbol{f}^{x_i} (i = 1, 2, \cdots, n)$,
Output: Set of seed examples $\boldsymbol{f}^{x_{s_j}} (j = 1, 2, \cdots, u)$
Method:
1. Initialize the first seed cluster using the input with the maximal norm, i.e., $c_1 = \boldsymbol{f}^{x_{J1}} \equiv argmax \|\boldsymbol{f}^{x_j}\|$
2. For $i = 2, \cdots, K$, each c_i is initialized in the following way: for each input features of examples $\boldsymbol{f}^{x_j}$, calculate its distance to the closest seed cluster
   $d_j = \min\{\|\boldsymbol{f}^{x_j} - c_k\|$:for all existing $c_k\}$,
   and set $c_i = \boldsymbol{f}^{x_{Ji}} \equiv argmax\{d_j\}$

図 2　KKZ アルゴリズム
Fig. 2　KKZ algorithm.

Algorithm: Introducing "must-link" constraint
Input: Clusters C_i and C_j, and corresponding centroids $\boldsymbol{G}_i, \boldsymbol{G}_j$.
Method:
1. Compute new centroid $\boldsymbol{G}^{new}$ between clusters C_i and C_j to be merged.
2. Compute the distance $D(\boldsymbol{G}^{new}, \boldsymbol{G}_i)$ between $\boldsymbol{G}^{new}$ and $\boldsymbol{G}_i$, and the distance $D(\boldsymbol{G}^{new}, \boldsymbol{G}_j)$ between $\boldsymbol{G}^{new}$ and $\boldsymbol{G}_j$.
3. If $D(\boldsymbol{G}^{new}, \boldsymbol{G}_i) < Th_{dis}$, and $D(\boldsymbol{G}^{new}, \boldsymbol{G}_j) < Th_{dis}$, then "must-link" constraints are introduced between C_i and C_j.

図 3　制約 "must-link" の導入法
Fig. 3　Method for introducing "must-link" constraint.

3.3　提案する半教師有りクラスタリング

本節では，我々が提案する半教師有りクラスタリングについて述べる．まず，クラスタの重心ベクトル $\boldsymbol{G}$ を式 (2) のように定義する．

$$\boldsymbol{G} = (g_1, g_2, \cdots, g_n) \tag{2}$$

ここで，g_k $(k = 1, 2, \cdots, n)$ はクラスタの重心ベクトルにおける各素性を表す．

我々の提案する手法では，この重心を計算する際，あるクラスタを種用例を含むクラスタにマージする場合に，そのクラスタの重心の変動を抑える点において，新規性がある．具体的には，種用例を含むクラスタの重心を再計算する場合に，式 (3) を用いる．この処理において，あるクラスタを種用例を含むクラスタにマージする際，そのクラスタの重心 $\boldsymbol{G}$ と $\boldsymbol{f}^x$ 間の距離 $D(\boldsymbol{G}, \boldsymbol{f}^x)$ によって，$\boldsymbol{f}^x$ の各要素 f_i^x $(i = 1, 2, \cdots, n)$ を重み付けする．本研究では，この距離尺度として，適応的マハラノビス距離を用いた．この距離尺度は，あるクラスタに属する用例数が少ないときに共分散が大きくなるというマハラノビス距離の問題点を解決する距離尺度であり，[8] において有効であることを確認している．

$$\boldsymbol{G}^{new} = \frac{\sum_{\boldsymbol{f}^x \in C_{(s)_i}} \boldsymbol{f}^x + \sum_{\boldsymbol{f}^x \in C_j} \frac{\boldsymbol{f}^x}{D(\boldsymbol{G}, \boldsymbol{f}^x) + c}}{n_{C_{(s)_i}} + n_{C_j}} \tag{3}$$

ここで，$n_{C_{(s)_i}}$ と n_{C_j} はそれぞれ，種用例を含むクラスタにおける用例数，ならびに種用例を含まないクラスタにおける用例数を表す．さらに，c は $D(\boldsymbol{G}, \boldsymbol{f}^x)$ が 0 に非常に近い値となったとき，$\boldsymbol{f}^x$ の各要素が極端に大きな値となることを防ぐために導入した定数であり，予備実験の結果，$c = 0.992$ と定めた．

一方，種用例を含まないクラスタどうしの重心の計算には，式 (4) を用いる．

$$\boldsymbol{G}^{new} = \frac{\sum_{\boldsymbol{f}^x \in C_i} \boldsymbol{f}^x + \sum_{\boldsymbol{f}^x \in C_j} \boldsymbol{f}^x}{n_{C_i} + n_{C_j}} \tag{4}$$

Algorithm: Semi-supervised clustering
Input: Set of features of examples $\boldsymbol{f}^{x_i}(i=1,2,\cdots,n)$,
and seed examples $\boldsymbol{f}^{s_j}(j=1,2,\cdots,u)$,
$E=\{\boldsymbol{f}^{x_1},\boldsymbol{f}^{x_2},\cdots,\boldsymbol{f}^{x_n},\boldsymbol{f}^{s_1},\boldsymbol{f}^{s_2},\cdots,\boldsymbol{f}^{s_u}\}$.
Output: Set of clusters $\mathcal{C}=\{C_1,C_2,\cdots\}$ that contain the examples that have the same sense.
Method:
1. Set feature of each example $\boldsymbol{f}^{x_i}$
and each feature of seed example $\boldsymbol{f}^{s_j}$
as an initial cluster C_i and $C_{(s)_j}$, respectively.
$C_i=\{\boldsymbol{f}^{x_i}\}, C_{(s)_j}=\{\boldsymbol{f}^{s_j}\}$,
thus, set of clusters $\mathcal{C}=\{C_1,C_2,\cdots,C_n,C_{(s)_1}\cdots C_{(s)_u}\}$,
where constraints are introduced
between examples $C_{(s)_m}$ and $C_{(s)_n}$.
2. **do**
2.1 Compute the similarity between C_i and C_j, C_i and $C_{(s)_k}$.
if the maximum similarity is obtained
between C_i and $C_{(s)_k}$, and the similarity $> Th_{sim}$
then merge C_i into $C_{(s)_k}$ to form a new cluster C^{new},
add C^{new} to $\mathcal{C}$, remove C_i from $\mathcal{C}$,
and recompute the centroid of
the cluster using Equation (3),
else if the maximum similarity is obtained
between C_i and C_j, and the similarity $> Th_{sim}$
then merge C_i and C_j to form a new cluster C^{new},
add C^{new} to $\mathcal{C}$, remove C_i and C_j from $\mathcal{C}$,
and recompute the centroid of
the cluster using Equation (4).
2.2 Compute similarities between C^{new}
and all $C_i \in \mathcal{C}, C_i \neq C^{new}$
3. **until** All of the similarities between C_i and C_j are
less than the predefined threshold.
4. **return** Set of clusters $\mathcal{C}$.

図 4　提案する半教師有りクラスタリングアルゴリズム
Fig. 4　Our proposed semi-supervised clustering algorithm.

表 1　語義数と対応する対象語
Table 1　Numbers of word senses and their corresponding target word.

語義数	対象語
5	「社会」,「目」,「見る」,「受ける」
6	「近く」,「手」,「開く」,「進む」
7 以上	「間」,「頭」,「もの」,「かかる」,「出す」,「出る」,「取る」,「入る」,「持つ」

ここで，n_{C_i}，n_{C_j} は，それぞれ，クラスタ C_i，C_j に含まれる単語の用例数を表す．図 4 に，我々の提案する半教師有りクラスタリングの基本アルゴリズムを示す．

4. 実　　験

4.1 実験データ

本研究では，「SENSEVAL-2 日本語辞書タスク」で配布された RWC コーパスを用いた．このコーパスは，毎日新聞の 1994 年の 3,000 記事に対して，人手で語義タグが付与されている．語義タグは，品詞が名詞，動詞，形容詞のいずれかであり，岩波国語辞典に見出しのある，多義の単語（総計 148,558 語）に付与されている．

また，「SENSEVAL-2 日本語辞書タスク」で用いられた 100 単語（名詞，動詞がそれぞれ 50 単語）のうち，語義数が 5 以上の単語（名詞 7 語，動詞 10 語）を対象語とした．これは，半教師有りクラスタリングにおいて，複数の種用例を導入することによる効果を確認するためである．表 1 に，本研究で用いた対象語を示す．

4.2 評価尺度

本研究では，“purity”，“inverse purity” と，これらの調和平均である F 値に基づいて，クラスタリングの精度を評価する．以下，生成されたクラスタに割り当てられるべき，人手で定めた正解を「カテゴリ」と呼ぶことにする．“purity” は「適合率」の尺度に関連する．この尺度では，各クラスタにおいて最もよく現れるカテゴリの頻度に注目し，ノイズの少ないクラスタを高く評価する．C を評価されるべきクラスタの集合，L を人手で作成したカテゴリの集合，n を生成されたクラスタ数とすると，purity は，式 (5) に基づいて，最大となる適合率の重み付き平均をとることで計算される．

$$Purity=\sum_i \frac{|C_i|}{n}\max Precision(C_i,L_j) \qquad (5)$$

ここで，ある与えられたカテゴリ L_j に対するクラスタ C_i の適合率 $Precision(C_i,L_j)$ は，式 (6) によって定義される．

$$Precision(C_i,L_j)=\frac{|C_i\bigcap L_j|}{|C_i|} \qquad (6)$$

“inverse purity” は，各カテゴリに対して最大の再現率となるクラスタに着目する．ある一つのクラスタにおいて，各カテゴリで定められた要素が多く集まったクラスタを高く評価する．inverse purity は，式 (7) によって定義される．

$$InversePurity=\sum_j \frac{|L_j|}{n}\max Recall(C_i,L_j) \qquad (7)$$

ここで，ある与えられたカテゴリ L_j に対するクラスタ C_i の再現率 $Recall(C_i,L_j)$ は，式 (8) によって定義される．

$$Recall(C_i,L_j)=\frac{|C_i\bigcap L_j|}{|L_j|} \qquad (8)$$

また，purity と inverse purity の調和平均 F は，式 (9) によって定義される．

$$F=\frac{1}{\alpha\frac{1}{Purity}+(1-\alpha)\frac{1}{InversePurity}} \qquad (9)$$

なお、本研究では，$\alpha=0.5$ として，評価を行なった．

4.3 実験結果

図 5 に手法 I における，表 1 の各語義数についてのクラスタリング精度を示す．また，紙面の都合上，手法 II，III については，最も良いクラスタリング精度が得られた制約 (d) を導入した場合についての実験結果のみを図 6，7 に示す．

4.4 考　　察

図 5，6，7 から，全体的な傾向として，**3.2** で述べた種用例の導入法，制約の導入法によらず，重心の変動を抑えたクラスタリングを行なう提案手法が，もっとも良いクラスタリング精度を示した．一方，距離を学習するクラスタリング手法においては，Bar-Hillel ら [6]，Xing ら [5]，Klein ら [4] の手法の順に良いクラスタリング精度が得られている．Klein らの手法では，類似した 2 点 (x_i,x_j) 間を 0，類似していない 2 点間を $(\max_{i,j} D_{ij})+1$ と設定した単純な隣接行列を作成した上で，クラスタリングを行なうのに対し，Xing ら，Bar-Hillel らの方法では，特徴空間を適切に変換する手法が用いられている．後者の 2 手法では，この変換手法が有効に作用しているものと考えられる．しかし，これらの距離を学習する手法と比較しても，重心の変動を抑えたクラスタリングを行なう我々の提案手法が，最も良いクラスタリング精度を示した．これは，空間を大域的に変換することになる距離を学習するクラスタリン

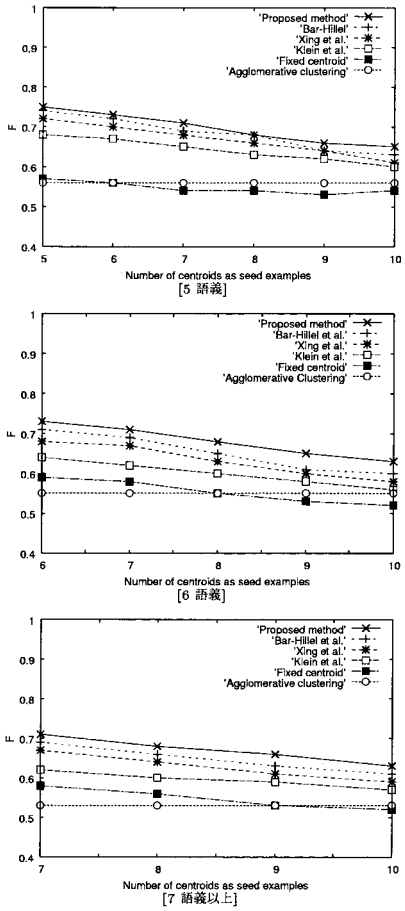

図 5　手法 I によるクラスタリング精度
Fig. 5　Clustering accuracy obtained using Method I.

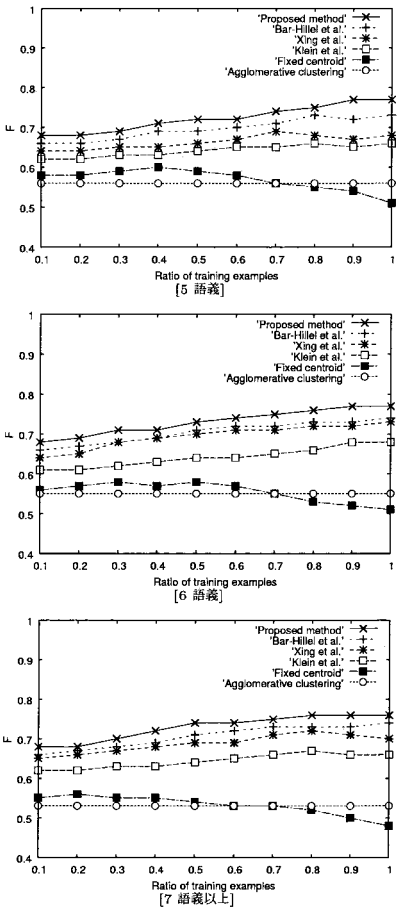

図 6　手法 II による種用例の導入法，ならびに (d) による制約を導入した場合の，クラスタリング精度
Fig. 6　Clustering accuracy obtained using Method II and (d) for introducing seed examples and constraints, respectively.

グ手法よりも，空間内のある点をクラスタにマージするたびに，局所的に重心を調整していくことによる効果であると考えられる．

また，重心を固定してクラスタリングを行った場合，その重心と類似度が高い用例しかマージされなくなるため，本来クラスタにマージすべき用例が独立したクラスタとなってしまう，すなわち，purity が大きく，inverse purity が小さくなる傾向が観察される．また，種用例が多くなるほど，独立したクラスタが多くなるため，その傾向は強まることが観察された．

さらに，凝集型クラスタリングは，制約によらないために精度は一定となる．用例数が多い場合に，重心を固定する方式よりも精度が良いのは，マージする用例に応じて，重心が変動するためであると考えられる．

また，**3.2** で述べた種用例の導入法に関して，手法 I による種用例の導入法では，その語義数に相当する個数の種用例を導入したときに，いちばん良い精度が得られた．これは，他の語義数のところでは，purity は中程度の値が得られているが，inverse purity の値が悪くなっており，その結果として低い F 値が得られたことによる．また，7 語義以上の対象語の場合，5 語義，6 語義の対象語と比較して，種用例数が増えるにつれ，クラスタリング精度がそれほど減少していないのは，7 語義以上の場合には，語義数が混在しているため，これらの語についてのクラスタリング精度が高いことによると考えられる．

一方，手法 II，手法 III では，制約 (d) を導入することで，最も良いクラスタリング精度が得られた．これは，例外値を除くことで，クラスタの重心の正確性が増したためであると考えられる．また，手法 II，III ともに，種用例を選択する訓練用例の割合が増えるにつれ，緩やかにクラスタリング精度が上昇する傾向が観察された．しかしながら，種用例を選択する訓練用例の割合が最も大きい場合に，手法 II，手法 III で得られたクラスタリング精度には，大きな差がない．したがって，[11] で述べられているように，精度の高いクラスタリング結果が得られることは確認されたが，訓練用例すべてを種用例とする手法 II を用いたほうが，簡潔な手法でもあり，効果的であると考えられる．

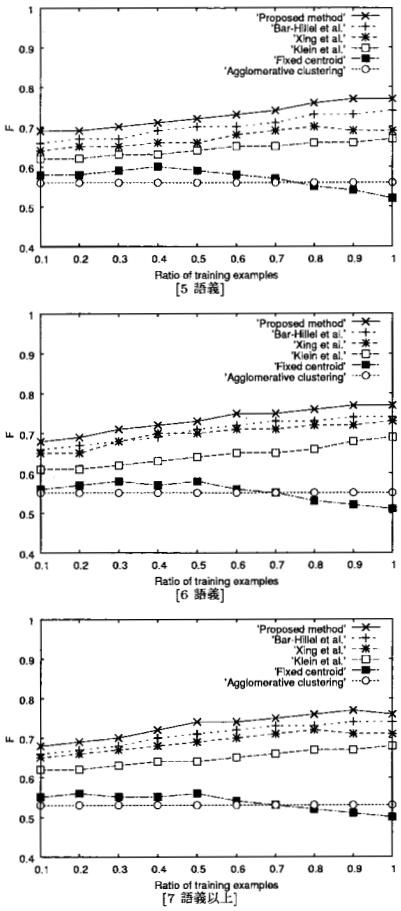

図 7 手法 III による種用例の導入法，ならびに (d) による制約を導入した場合の，クラスタリング精度

Fig. 7 Clustering accuracy obtained using Method III and (d) for introducing seed examples and constraints, respectively.

5. おわりに

本論文では，単語の用例のための半教師有りクラスタリングの手法について提案した．我々の手法は，種用例を含むクラスタの重心の変動を抑えることによって，より正確な半教師有りクラスタリングを実現する点，クラスタが生成されてきた段階で，“must-link” の制約を導入することで，クラスタリングの精度を向上させることができた．しかし，この “must-link” の導入の仕方には，もう少し改善の余地がある．例えば，マージ後にクラスタ内分散を大きくしてしまうようなクラスタには，“must-link” を導入しないという方法も考えられる．こうした制約の導入法について，さらに検討を進めていく予定である．

文　献

[1] K. Wagstaff and C. Cardie. Clustering with Instance-level Constraints. In *Proc. of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1103–1110, 2000.

[2] K. Wagstaff and S. Rogers and S. Schroedl. Constrained K-means Clustering with Background Knowledge. In *Proc. of the 18th International Conference on Machine Learning (ICML 2001)*, pp. 577–584, 2001.

[3] S. Basu and A. Banerjee and R. Mooney. Semi-supervised Clustering by Seeding. In *Proc. of the 19th International Conference on Machine Learning (ICML 2002)*, pp. 27–34, 2002.

[4] D. Klein and S. D. Kamvar and C. D. Manning. From Instance-level Constraints to Space-level Constraints: Making the Most of Prior Knowledge in Data Clustering. In *Proc. of the 19th International Conference on Machine Learning (ICML 2002)*, pp. 307–314, 2002.

[5] E. P. Xing and A. Y. Ng and M. I. Jordan and S. J. Russell. Distance Metric Learning with Application to Clustering with Side-Information. *Advances in Neural Information Processing Systems*, Vol. 15, pp. 521–528, 2003.

[6] A. Bar-Hillel and T. Hertz and N. Shental. Learning Distance Functions Using Equivalence Relations. In *Proc. of the 20th International Conference on Machine Learning (ICML 2003)*, pp. 577–584, 2003.

[7] N. Shental and T. Hertz and D. Weinshall and M. Pavel. Adjustment Learning and Relevant Component Analysis. In *Proc. of the 7th European Conference on Computer Vision (ECCV 2002)*, pp. 776–792, 2002.

[8] 杉山一成，奥村学．Web 検索結果における人名の曖昧性解消への半教師有りクラスタリングの適用．情報処理学会研究報告 Vol.2007, No.94, 2007-NL-181 (3), pp. 15–20, 2007.

[9] J. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In *Proc. of the 5th Berkeley Symposium on Mathmatical Statistics and Probability*, pp. 281–297, 1967.

[10] I. Katsavounidis and C. Kuo and Z. Zhang. A New Initialization Technique for Generalized Lloyd Iteration. *IEEE Signal Processing Letters*, Vol. 1, No. 10, pp. 144–146, 1994.

[11] J. He and M. Lan and C.-L. Tan and S.-Y. Sung H.-B. Low. Initialization of Cluster Refinement Algorithms: A Review and Comparative Study. In *Proc. of the IEEE International Conference on Newral Networks*, pp. 297–302, 2004.

[12] H. Yang and J. Callan. Near-Duplicate Detection by Instance-level Constrained Clustering. In *Proc. of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pp. 421–428, 2006.

[13] W. Tang and H. Xiong and S. Zhong and J. Wu. Enhancing Semi-Supervised Clustering: A Feature Projection Perspective. In *Proc. of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD'07)*, pp. 707–716, 2007.

[14] I. Davidson and S. S. Ravi. Intractability and Clustering with Constraints. In *Proc. of the 24th International Conference on Machine Learning (ICML 2007)*, pp. 201–208, 2007.