用例のクラスタリング結果を利用した 語義曖昧性解消手法

杉山 一成† 奥村 学‡
†‡ 東京工業大学 精密工学研究所
† sugiyama@lr.pi.titech.ac.jp ‡ oku@pi.titech.ac.jp

1 はじめに

語義曖昧性解消は，ある文脈において現れる単語の意味を決定する処理であり，要約，質問応答，機械翻訳，情報検索などの自然言語処理の応用分野において有用である．この語義曖昧性解消のための手法として，教師有りの手法 [1]，半教師有りの手法 [2]，教師無しの手法 [3] が，これまでに提案されている．教師有りの手法は，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することができる．一方，教師無しの手法は，人手でタグ付けされたデータは不要であり，対象語が用いられている文脈間の類似性や単語の共起情報などを利用することにより，語義の曖昧性を解消する．さらに，半教師有りの手法は，ブートストラップの手法に基づいて，語義タグを付与した少数の種となるデータから，事例全体を分類するごとに信頼性のある訓練事例を加えていくことで，語義の曖昧性を解消する．これらの手法のうち，教師有りの語義曖昧性解消手法は，上述したように，高い精度を実現することができる一方，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グループ化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，用例をグループ化する際には，グループ間に制約を与えることで，各グループ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．そこで，語義タグ付きコーパスの作成を支援するために，我々は，制約情報を利用するクラスタリング手法である半教師有りクラスタリングに着目している．また，このクラスタリング結果を，教師有りの語義曖昧性解消における分類器を構築するための素性として活用すれば，その精度を改善することができるものと考えられる．そこで，本論文では，このクラスタリング結果を利用した素性を用いて，語義曖昧性を解消する手法について提案し，クラスタを考慮した素性が有効であることを示す．

2 関連研究

最近の語義曖昧性解消についての研究には，次のような研究を挙げることができる．

教師有りの手法について，Specia ら [4] は，素性ベクトルを作成して機械学習を用いた場合，限定された知識しか用いることができないことに着目し，素性表現をより豊かにするために，帰納論理プログラミングを適用して，動詞の語義曖昧性解消を行なった．その結果，多様なルールを生成する帰納論理プログラミングは，語義曖昧性解消には有効であるとの結論を得ている．また，Cai ら [5] は，まず，LDA (Latent Dirichlet Allocation) を用いて，20 Newsgroup などの語義タグ付けされていないコーパス中の語をクラスタリングすることによって，トピックモデルを構築する．次に，得られたトピックモデルを使って，語義タグ付けされたコーパスにおける単語の素性とする．このようにして得られた素性とともに，単語の表記，品詞，構文，連語から得られた素性を用いて，Support Vector Machine によって分類器を構築し，教師有りの語義曖昧性解消を実現している．

半教師有りの手法に関して，Niu ら [6] は，類似した用例は，類似した語義タグを持つはずであるという仮定に基づいて，学習過程において，語義タグ付けされたデータと語義タグ付けされていないデータを結びつける語義タグ伝播法を用いた半教師有り学習による語義曖昧性解消手法を提案している．

また，教師無しの手法において，Boyd-Graber ら [7] は，まず，WordNet の階層情報を隠れ変数として変換し，LDA を用いてトピックモデルを構築した後，事後確率を推論する問題として，語義曖昧性解消をとらえている．Brody ら [8] は，これまでに提案されている教師無しの語義曖昧性解消の手法のうち，精度の良い手法を組み合わせて語義曖昧性解消を行なう手法について提案し，単独の手法よりも良い精度が得られることを示している．

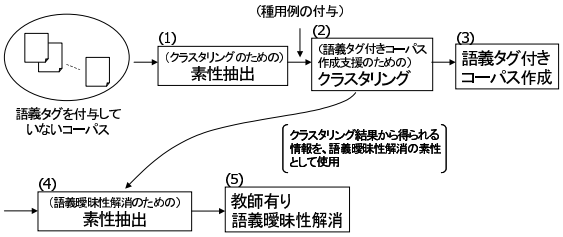

図 1: 語義曖昧性解消手法の構成

3 提案手法

3.1 提案手法の構成

図 1 に我々の提案する語義曖昧性解消手法の構成を示す．(1) 語義タグを付与していないコーパスから，クラスタリングのための素性を抽出する．この時に用いる素性については，3.2.1 節で述べる．(2) この抽出した素性を用いて，クラスタリングを行なう．このクラスタリング時には，複数の種用例と，その種用例間の関係を考慮した制約を導入する．これらの導入法については，3.2.2 節において述べる．また，このクラスタリングの処理においては，重心の変動を抑えることに着目した半教師有りクラスタリングを適用する．(3) 語義タグ付きコーパスの作成は，生成された各クラスタにおいて，最も多く出現する種用例の語義を，クラスタ内の各用例に付与する．(4) 語義タグを付与したコーパスから，語義曖昧性解消を行なうための素性を抽出するとともに，(2) のクラスタリング結果から得られる情報も素性として抽出し，(5) 教師有りの語義曖昧性解消を行なう．

3.2 クラスタリング手法の概要

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 [9]，[10]，[11]，(2) 距離に基づいた手法 [12]，[13]，[14] の二つに分類することができる．しかし，これらの手法は，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おり，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，半教師有りクラスタリングにおいては，より正確なクラスタリング結果を得るためには，用例間に制約を導入するとともに，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．そこで，3.1

節で述べたように，重心の変動を抑えることに着目した半教師有りクラスタリングを適用する [15].

3.2.1 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 (+2，+1) にある単語の位置を表す．なお，形態素解析器としては ChaSen[1] を，構文解析器としては CaboCha[2] を用いた．なお，以下において，これらの素性を “baseline” と呼ぶことにする．

3.2.2 種用例，ならびに制約の導入法

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．

[手法 I] 訓練事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 [16] を用いた．

[手法 II] 訓練事例すべてを種用例とする．

[手法 III] 訓練事例に “KKZ” [17] と呼ばれる手法を適用し，種用例を生成する．

手法 III は，クラスタの初期化法について比較した [18] において，精度の高い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おり，本研究においても有効であるかを確認するために導入した．

また，これらの 3 つの種用例の導入法において，次のように制約を導入する．まず，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．

(a) 異なる語義どうしに “cannot-link” を導入

(b) 同じ語義どうしに “must-link” を導入

(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

(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で，クラスタ間類似度が設定した閾値よりも大きくなった場合に，“must-link” を導入する．

なお，(d) においては，“must-link” 導入時に，例外的な種用例は除くようにし，クラスタ内分散の値が大きくなることを防ぐようにしている．

[1] http://sourceforge.net/projects/masayu-a/
[2] http://sourceforge.net/projects/cabocha/

3.3 語義曖昧性解消手法

本研究では，教師有りの語義曖昧性解消を実現するために，この分野の研究においてしばしば用いられる，Support Vector Machine (以下，SVM) [19], Naïve Bayes (以下，NB) [20], Maximum Entropy (以下，ME) [21] の機械学習手法を適用することにより，分類器を構築する．このために，本研究では，3.2.1 節で述べた “baseline” の素性のほかに，生成された各クラスタにおいて計算した次の素性を，分類器を構築するために導入する．すべての用例に対して，これらの素性を計算するのではなく，類似した用例を集めたクラスタ内について，これらの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ものと期待される．

(1) 単語の出現頻度 (以下，TF)
  - 特に，対象語の周囲によく現れる単語の頻度が大きい値となれば，対象語の語義を特定しやすくなるものと考えられる．なお，各クラスタによって単語の総数が異なるので，各クラスタ内の総単語数で正規化している．

(2) クラスタ ID (以下，CID)
  - 各クラスタには，類似した用例が集められている．したがって，その ID を素性として使用すれば，正解の語義を付与したことと，同等の効果が得られると考えられる．

(3) 種用例の語義の出現頻度 (以下，SF)
  - 3.2.2 節で述べたように，本研究では種用例を導入した半教師有りクラスタリングを行なう．したがって，そのクラスタに含まれる種用例の語義の出現頻度をクラスタ内の用例の素性として加えることで，対象語の語義を特定しやすくなるものと考えられる．

(4) 相互情報量 (以下，MI)

(5) 情報利得 (以下，IG)

(6) T-score (以下，T)

(7) χ^2 値 (以下，CHI2)
  - (4)〜(7) は，連語を抽出する際にしばしば用いられるスコアである．“one sense per collocation” [2] の概念を反映する指標として導入している．また，これらは，$W(-2)W(-1)$, $W(-1)W(0)$, $W(0)W(+1)$, $W(+1)W(+2)$ の隣接する 2 単語間について計算する．

(8) 各クラスタにおいて，種用例以外の用例をグループ化した ID (以下，GrpID)
  - 各クラスタ内における種用例以外の用例について，凝集型クラスタリングを行ない，生成されたクラスタの ID を付与する．クラスタ内において，さらに用例をまとめることで，さらに語義を特定しやすくなるものと考えられる．

(9) (1)〜(8) のすべて (ALL)

例えば，図 2 のようなクラスタが構築された場合，上述した素性は，次のように計算される．まず，クラスタ C_1 において，種用例 $f_{s_2}(2)$，$f_{s_4}(2)$，$f_{s_7}(4)$，種用例以外の用例 f_1, f_3, f_5, f_9, について，$W(-2)$〜$W(+2)$ の各単語の頻度を計算し，各クラスタ内の総単語数で正規化する．また，これらの 7 つの各用例には，クラスタ ID である C_1 から “1” を付与する．語義の出現頻度の素性については，語義 ID “2” の種用例が 2 つ，語義 ID “4” の種用例が 1 つ含まれるので，このクラスタ C_1 内の各用例に，語義の出現頻度の値 “2”，“1” を付与する．また，クラスタ C_1 に出現している隣接 2 単語間について，(4)〜(7) のスコアを計算する．さらに，C_1 において，種用例以外の用例について，凝集型クラスタリングによって，f_3，f_5 が同じクラスタ g_2 となり，f_1，f_9 がそれぞれ独立したクラスタ g_1，g_3 にグループ化されたとすると，f_1 には “1” を，f_3，f_5 には “2” を，f_9 には “3” の値を付与する．また，種用例以外の用例には，そのクラスタにおいて，最も多い種用例の語義を付与する．すなわち，クラスタ C_1 において

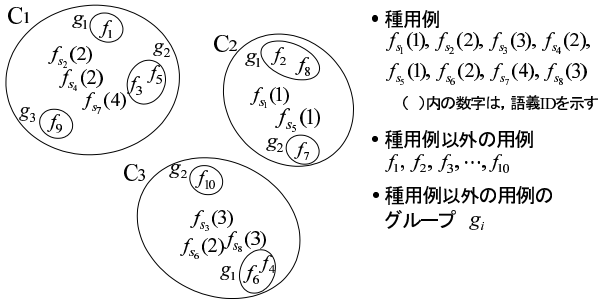

図 2: ある対象語の用例について生成されたクラスタ

表 1: 語義数と対応する対象語

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」

は，f_1，f_3，f_5，f_9 に，語義 ID“2” を付与する．クラスタ C_2，C_3 についても，上述したクラスタ C_1 の処理と同様の処理を行なう．以上のようにして，ある対象語についての教師有りデータを作成した後，5 分割交差検定を行ない，語義曖昧性解消の精度を検証した．

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．

本研究では,「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，複数の種事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

用例のクラスタリング結果の評価には，“purity，” “inverse purity” の調和平均を計算した F 値を用いた．

また，語義曖昧性解消の評価尺度としては，精度を用いた．この精度は，正解の語義 ID とシステムの出力する語義 ID が完全に一致していれば正解とする “fine-grained scoring” [22] を評価基準としたものである．

4.3 実験結果

まず，用例のクラスタリングについての評価結果を，表 2 に示す．手法 I においては，語義数相当の種用例を生成した場合に，手法 II，III においては，全訓練データを用い，3.2.2 で述べた (d) の制約を導入した場合に，最も良いクラスタリング精度が得られた．

また，この最も良いクラスタリング精度が得られた場合に，3.3 節で述べた素性を “baseine” の素性に追加して，分類器を構築した．SVM，NB，ME の各機械学習手法を用いて得られた語義曖昧性解消の精度を，それぞれ表 3，4，5 に示す．これらの表において，“baseline” は 3.2.1 節で述べた素性を導入した結果である．また，

表 2: クラスタリング精度

語義数	手法 I [語義数相当のクラスタ 生成＋“cannot-link”]	手法 II [全訓練データ ＋(d) の制約]	手法 III [全訓練データ ＋(d) の制約]
5	0.75	0.76	0.77
6	0.73	0.75	0.77
7 以上	0.71	0.74	0.76

表 3: SVM による語義曖昧性解消の精度

[用例のクラスタリングなし]

語義数	baseline	baseline +MI	baseline +IG	baseline +T	baseline +CHI2
5	0.744	0.747	**0.748**	0.746	0.745
6	0.665	0.668	**0.670**	0.667	0.668
7 以上	0.656	**0.659**	0.657	0.657	0.656

[手法 I (語義数相当のクラスタ生成＋“cannot-link”)]

語義数	baseline	baseline +SF	baseline +IG	baseline +GrpID	ALL
5	0.744	0.750	0.748	0.747	**0.752**
6	0.665	**0.672**	0.671	0.669	**0.672**
7 以上	0.656	0.661	0.659	0.658	**0.663**

[手法 II (全訓練データ+(d) の制約)]

語義数	baseline	baseline +SF	baseline +IG	baseline +GrpID	ALL
5	0.744	0.748	0.748	0.746	**0.750**
6	0.665	0.670	0.670	0.667	**0.671**
7 以上	0.656	0.660	0.659	0.657	**0.662**

[手法 III (全訓練データ+(d) の制約)]

語義数	baseline	baseline +SF	baseline +IG	baseline +GrpID	ALL
5	0.744	0.752	0.748	0.746	**0.754**
6	0.665	0.673	0.670	0.667	**0.675**
7 以上	0.656	0.662	0.657	0.657	**0.665**

「手法 I」,「手法 II」,「手法 III」は，3.2.2 で述べた種用例の導入法を表す．なお，これらの表では，“baseline” の素性を用いて得られた結果と比較して，改善の程度が大きかった素性についての結果のみを示している．

4.4 考察

本節では，SVM，NB，ME の各学習法ごとに，用例のクラスタリングを行なった場合と行なわない場合のそれぞれについて，素性を加えることで得られた効果について考察する．

まず，表 3 から，SVM については，用例のクラスタリングを行なわない場合，“baseline” の素性に，IG を加えることで良い精度が得られた．また，クラスタリングを行なった場合には，手法 III の方法で種用例を導入し，すべての素性 (ALL) を導入した場合に，最も良い精度が得られた．しかし，用例のクラスタリングを行ない，そのクラスタ内で計算した素性を導入しても，クラスタリングを行なわない場合と比べて，それほど精度は改善されなかった．

次に，表 4 から，NB については，用例のクラスタリングを行なわない場合には，SVM と同様に，“baseline” の素性に，IG を加えることで，2〜4%程度，精度が改善された．また，用例のクラスタリングを行なった場合には，SF を素性として用いた場合，6 語義と 7 語義以上の単語において，“baseline” と比較して，3〜5%程度，精度が改善された．また，SVM と同様に，手法 III の方法で種用例を導入した場合に，最も良い精度が得られた．

さらに，表 5 から，ME については，用例のクラスタリングを行なわない場合には，素性を追加することで，1%程度，精度が改善された．また，用例のクラスタリングを行なった場合には，手法 I，II，III の種用例を導入する手法によらず，特にすべての素性 (ALL) を導入した場合において，2%弱の精度の改善が観察された．

以上から，用例のクラスタリングを行なわないよりも，一度用例のクラスタリングを行ない，類似した用例を集めたクラスタ内の用例について，3.3 節で述べた素性を導入することは，語義の曖昧性解消に有効であるといえる．これは，用例のクラスタリングを一度行なったことで，語義を限定する効果が反映されたためであると考えられる．

表 4: NB による語義曖昧性解消の精度

[用例のクラスタリングなし]

語義数	baseline	baseline +MI	baseline +IG	baseline +T	baseline +CHI2
5	0.759	0.792	**0.799**	0.794	0.793
6	0.703	0.711	**0.728**	0.704	0.705
7 以上	0.573	0.595	**0.611**	0.576	0.575

[手法 I (語義数相当のクラスタ生成+“cannot-link”)]

語義数	baseline	baseline +SF	baseline +IG	baseline +GrpID	ALL
5	0.759	0.811	0.810	0.805	**0.812**
6	0.703	**0.731**	0.728	0.709	0.730
7 以上	0.573	**0.621**	0.613	0.609	0.614

[手法 II (全訓練データ+(d) の制約)]

語義数	baseline	baseline +SF	baseline +IG	baseline +GrpID	ALL
5	0.759	0.812	0.812	0.810	**0.814**
6	0.703	**0.732**	0.730	0.710	0.730
7 以上	0.573	**0.623**	0.615	0.611	0.617

[手法 III (全訓練データ+(d) の制約)]

語義数	baseline	baseline +SF	baseline +IG	baseline +GrpID	ALL
5	0.759	0.813	0.814	0.810	**0.815**
6	0.703	**0.735**	0.732	0.712	0.730
7 以上	0.573	**0.624**	0.617	0.611	0.620

表 5: ME による語義曖昧性解消の精度

[用例のクラスタリングなし]

語義数	baseline	baseline +MI	baseline +IG	baseline +T	baseline +CHI2
5	0.780	0.785	**0.793**	0.784	0.786
6	0.610	0.617	**0.620**	0.614	0.613
7 以上	0.607	0.614	**0.617**	0.616	0.616

[手法 I (語義数相当のクラスタ生成+“cannot-link”)]

語義数	baseline	baseline +SF	baseline +IG	baseline +GrpID	ALL
5	0.780	0.788	0.795	0.785	**0.797**
6	0.610	0.618	0.621	0.616	**0.625**
7 以上	0.607	0.619	0.619	0.617	**0.621**

[手法 II (全訓練データ+(d) の制約)]

語義数	baseline	baseline +SF	baseline +IG	baseline +GrpID	ALL
5	0.780	0.790	0.797	0.788	**0.799**
6	0.610	0.621	0.624	0.618	**0.625**
7 以上	0.607	0.620	0.623	0.619	**0.625**

[手法 III (全訓練データ+(d) の制約)]

語義数	baseline	baseline +SF	baseline +IG	baseline +GrpID	ALL
5	0.780	0.792	0.797	0.788	**0.805**
6	0.610	0.623	**0.626**	0.620	0.625
7 以上	0.607	0.622	**0.627**	0.620	0.626

5 おわりに

本論文では，用例のクラスタリング結果を利用した素性を用いて，語義曖昧性を解消する手法について提案した．実験の結果，用例のクラスタリングを行なわない場合，種用例の語義頻度，情報利得を素性として導入することにより，精度が効果的に改善されることがわかった．また，用例のクラスタリングを行ない，生成された各クラスタにおいて計算した素性を加えた場合には，さらに精度が改善されることも確認された．この結果から，用例のクラスタリング結果に基づいて計算した素性を，形態素素性や構文素性に追加したことは，語義曖昧性解消に対して有効な手法であると考えられる．

今後の課題として，特に 7 語義以上の単語，すなわち語義数が多い単語に対して，より高い精度が得られるように，分類器を構築するための素性を工夫することが挙げられる．具体的には，文献 [23] で提案されている連語性のスコアを改良し，その有効性を確認する予定である．

参考文献

[1] C. Leacock and G. A. Miller and M. Chodorow. Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, Vol. 24, No. 1, pp. 147–165, 1998.

[2] D. Yarowsky. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL 1995)*, pp. 189–196, 1995.

[3] H. Schütze. Automatic Word Sense Discrimination. *Computational Linguistics*, Vol. 24, No. 1, pp. 97–123, 1998.

[4] L. Specia, M. Stevenson, and M. G. V. Nunes. Learning Expressive Models for Word Sense Disambiguation. In *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pp. 41–48, 2007.

[5] J. F. Cai, W. S. Lee, and Y. W. Teh. Improving Word Sense Disambiguation Using Topic Features. In *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP 2007)*, pp. 1015–1023, 2007.

[6] Z.-Y. Niu, D.-H. Ji, and C. L. Tan. Word Sense Disambiguation Using Label Propagation Based Semi-Supervised Learning. In *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pp. 395–402, 2007.

[7] J. Boyd-Graber, D. Blei, and X. Zhu. A Topic Model for Word Sense Disambiguation. In *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP 2007)*, pp. 1024–1033, 2007.

[8] S. Brody, R. Navigli, and M. Lapata. Ensemble Methods for Unsupervised WSD. In *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pp. 97–104, 2007.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] K. Sugiyama and M. Okumura. Personal Name Disambiguation in Web Search Results Based on a Semi-Supervised Clustering Approach. In *Proc. of the 10th International Conference on Asian Digital Libraries (ICADL'07)*, pp. 250–256, 2007.

[16] 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.

[17] 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.

[18] 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.

[19] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[20] G. H. John and P. Langley. Estimating Continuous Distributions in Bayesian Classifiers. In *Proc. of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI'95)*, pp. 338–345, 1995.

[21] A. L. Berger and S. D. Pietra and V. J. D. Pietra. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, Vol. 22, No. 1, pp. 39–71, 1996.

[22] 白井 清昭. SENSEVAL-2 日本語辞書タスク. 自然言語処理, Vol. 2, pp. 3–24, 2003.

[23] J. Wermter and U. Hahn. Collocation Extraction Based on Modifiability Statistics. In *Proc. of the 20th International Conference on Computational Linguistics (COLING'04)*, pp. 980–986, 2004.