Vol.2009-DBS-148 No.6
Vol.2009-FI-95 No.6
2009/7/28

Web上の人名検索結果の同姓同名問題における二段階クラスタリングを用いた再現率向上

池田雅紀[†1]　小野　真吾[†1]　佐藤　一誠[†1]
吉田　稔[†2]　中川　裕志[†2]

教師なし学習によるクラスタリングに対して，半教師有り学習を適用する手法について提案する．クラスターの評価基準において，結果のクラスターにおける正解データの割合を表す適合率と正解データが結果のクラスターに含まれている割合を表す再現率が存在する．従来研究において，素性の種類を限定することによって特に高い適合率を持つクラスターを生成することが可能になった．これらの素性は疎であり，再現率を向上させることは困難である．一方，素性の中には，人物を識別する能力は弱いが，文書に含まれている数の多い素性が存在する．我々は半教師有り学習を適合率の高いクラスターに対して適用し，クラスターの再現率の向上させることを提案する．本研究では，ブートストラップ法として知られている Espresso を応用し，人名曖昧解消における半教師有り学習として用いる．

Improvement Recall of Person Name Disambiguation on the Web People Search by TwoStage Clustering

Masaki Ikeda [†1], Shingo Ono [†1], Issei SATO [†1]
Minoru YOSHIDA [†2] and Hiroshi Nakagawa [†2]

This research proposes the application of semi-supervised learning to unsupervsed clustering. There are two criteria of cluster evaluation, or precision and recall. Precision is the ratio of true datas in the result cluster and recall is the ratio of true datas the result cluster has to all true data. In previous work, the selection of feature types enables to make high precision clusters, but these features are too sparse to imporve recall. On the otherhand, there are features that has poor discrimination capacity but are thick in the documents. We suggest to applicate semi-supervised learning to these high precision clusters and advance clusters' recall. In this research, we use Espresso that is bootstrap method in the information extraction for person name disambiguation as semi-supervised learning method.

1. 導　　入

Web 上の人物検索は Web 検索において重要な地位を占めてきている．このような状況の中，人物の検索に関する問題として人物の同姓同名問題の解消が求められている．人物の同姓同名問題とは，Web 検索において検索対象者と同姓同名の人物の存在によって検索結果から目的の人物のページを発見することが困難になるという問題である．特に困難な場合としては以下の場合が考えられる．第一に，検索対象者と同姓同名の有名人が存在する場合である．例えば，米国前大統領の “George Bush” と同姓同名の別人物を検索する場合，大統領である “George Bush” に関するページが検索結果に多く現れ，目的とするページを探すのが困難になる．第二に，検索対象者の名前が多くの同姓同名の人物を持つ場合である．例えば，“田中太郎”，“John Smith” という名前を持つ人々は非常に多い．このように，同姓同名問題は言語を問わず問題となっている．この問題の解決方法として提案されているのが，検索結果の人名ごとのクラスタリングである．即ち，検索結果を同一人物ごとのクラスタにまとめて提示し，検索結果の閲覧性を向上させることで同姓同名の存在による効率の低下を防ぐという方法である．

同姓同名の人物のクラスタリングには文書中の人物に関わる名詞句を用いることが有効であるとされている．特に，人名，地名，組織名といった固有表現がクラスタリングにおいて有効であると先行研究[11)] によって示されている．しかし，これらの素性は極めて疎であるため，全ての文書間で類似度を計算することは難しく，再現率を向上させることは困難である．一方，単語などの素性は人物を識別する能力は弱いが，多くの文書に出現する．本研究では，疎な素性を用いて高い適合率を持つクラスターを作成した後，識別性能の低い密な素性から相対的に識別性能の高い素性を選択してクラスターを拡張し，再現率の向上を図る．この問題は初期クラスターに含まれる文書を labeled data とする半教師有り学習の問題とみなすことができる．本研究では半教師有り学習の手法の一種であるブートストラップを用いてクラスターの再現率の向上を目指す．自然言語処理におけるブートストラップとは

†1 東京大学大学院情報理工学系研究科 〒 113-0033 東京都文京区本郷 7-3-1
Graduate School of Information Science and Technology, The University of Tokyo. Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-0033 Japan
†2 東京大学情報基盤センター
Information Technology Center, The University of Tokyo

Vol.2009-DBS-148 No.6
Vol.2009-FI-95 No.6
2009/7/28

少数の labeled data を元に多数の unlabeled data にラベルを割り当て，反復的にラベルを割り振っていく手法である．本手法の評価は英語における同姓同名人物のクラスタリングタスクである WePS[6] のデータセットを用いて行う．

本稿の構成は，以下のようになっている．第 2 節では関連研究について述べる．第 3 節では高精度クラスターにおける再現率改善についての概要を述べる．第 4 節では再現率改善のために用いる，ブートストラップ手法について説明する．第 5 節では高精度クラスターの生成方法について説明する．第 6 節では実験により本研究の手法を検証した結果について説明する．第 7 節で本稿の結論を述べる．

2. 関 連 研 究

関連研究として，以下のようなものがある．Bagga ら[7] は，文書に出現する単語を要素とする文書ベクトルを作り，ベクトル空間内において文書間の類似度を計算し，クラスタリングを行った．出現する単語に加え，文書中から人物に関する個人情報を抽出し，クラスタリングする試みとして 文献 15) が挙げられる．

本研究で用いている二段階クラスタリングに関する先行研究として次のような研究が挙げられる．Bekkerman ら[9] は，実世界において同一コミュニティに属する複数の人物に関するページを集め，それらの中でのリンク解析や階層併合・分割ダブルクラスタリングを行うことで，同一コミュニティに属する人物を同時にそれぞれの同姓同名についての記述と分離する方法を提案した．この方法では結果的に同じコミュニティに属する人物について前提知識として情報が与えられた上でクラスタリングを行っていることになる．Tishby ら[22] によって提案された情報ボトルネック法は情報理論を用いて，最適なクラスタリングを求めるアルゴリズムである．情報ボトルネック法は Slonim ら[21] によって文書クラスタリングに対して適用されている．彼らは文書クラスタリングに対して，関連すると考えられる単語クラスタリングの結果を用いてクラスタリングを行っている．Liu ら[14] はクラスタを区別するために有効な特徴量を一段階のクラスタの多数決に基づいて求め，K-means を用いて，二段階クラスタリングを行っている．これらの手法は一段階目の結果クラスターを直接二段階目のクラスタリングには反映させていない．本研究では，二段階クラスタリングによって一段階目の結果クラスターを直接拡張している．

半教師有り学習の代表的な手法として，語義曖昧性解消における半教師有り学習である Yarowsky Algorithm[23] が知られている．Yarowsky Algorithm は語義のコロケーションにおける一貫性と文書内における同一性に基づいて，labeled data を増やしていく方法であり，decision list と呼ばれる対数尤度表を用いて，unlabeled data に対して，ラベルを割り当てている．この Yarowsky Algorithm について理論的な解析を行った論文として，Abney らの文献 1)，2) がある．また，情報抽出におけるブートストラップ手法として知られている Espresso[18] がある．この手法は同義語などのインスタンスとパターンを自己相互情報量に基づいて，抽出する手法である．Espresso を理論的な解析を行った論文として，Komachi らの文献 13) がある．本研究では，Espresso を応用し，自己相互情報量を用いて人名曖昧性解消のためのブートストラップを行う．

また，近年人名の曖昧性の解消を目的とした Web 上での人物検索に関するワークショップ WePS[4] が行われ，様々な知見が明らかとなっている．2006 年から 2007 年にかけて第 1 回が行われ，2008 年から 2009 年に第 2 回が行われた[6]．WePS の上位チーム[8),10),11),19),20)] が用いている方法の多くは文書ベクトル空間の類似度に基づくクラスタリングを用いたものである．本手法のように，二段階クラスタリングによってクラスターを拡張する手法はとられてはいない．

3. 半教師有り学習に基づく再現率の改善

ここでは本研究における提案手法であるブートストラップを用いて，既存クラスターの再現率を改善する手法について説明する．

3.1 適合率・再現率

人物についてのクラスタリングについて，情報検索のシステムの評価に用いられる適合率・再現率の観点から説明する．同姓同名人物のクラスタリングの目的は同姓同名人物を区別し，同一人物ごとのクラスターを作成することである．このとき，評価対象となる結果のクラスター集合を $\mathcal{C} = \{C_1, \cdots, C_i, \cdots, C_N\}$ とし，人手で作成した同一人物ごとのクラスター集合を正解のクラスター集合として $\mathcal{L} = \{L_1, \cdots, L_j, \cdots, L_M\}$ とする．

クラスター C_i に対して，ある基準で正解集合 $\mathcal{L}$ の中から正解のクラスター L_j を選ぶ．このクラスター L_j に含まれる文書を正解文書とする．

適合率 (Presicion) とは生成されたクラスターにおいて，正解文書がどれだけ含まれているかを表す指標である．再現率 (Recall) とは生成されたクラスターが，正解の文書集合のうちどれだけを含んでいるかを表す指標である．生成された 1 クラスターを対象として，適合率・再現率の計算方法について説明する．適合率 P・再現率 R は式 (1),(2) によって表わすことができる．

Vol.2009-DBS-148 No.6
Vol.2009-FI-95 No.6
2009/7/28

$$P = \frac{|C_i \cap L_j|}{|C_i|} \tag{1}$$

$$R = \frac{|C_i \cap L_j|}{|L_j|} \tag{2}$$

3.2 素性によるクラスターの違い

Web 上の文書を対象にして，人物についてのクラスタリングを行う場合，素性として，住所やメールアドレスなどの個人情報，その人物に関わる人名や組織名（固有表現），専門用語や単語 (N-gram) などが考えられる．また，Web 上の文書を対象にしていることからリンク先やリンク元の URL も素性となる．

適合率・再現率の観点からこれらの素性について考えると，人物との関連度の違いによって，生成されるクラスターの適合率・再現率が異なる．人物との関連度が強い素性として，個人情報や固有表現が挙げられる．これらの素性を用いることで人物を識別し，同一人物について参照を行っている文書を高い確率でまとめることができる．しかし，このような識別性能の高い素性は文書中にあまり出現せず，もう一つの指標である再現率を向上させることは難しい．一方，人物との関連度が弱い素性として，単語などが挙げられる．これらの素性は文書中で出現する頻度は多く，対象人物を参照していると考えられる文書を多く発見することが可能である．しかし，全体的に識別性能が低く，誤ったクラスターを形成しやすい．そのため，生成されたクラスターは再現率が高く，適合率が低いものになりやすい．このように，人物との関連性の強弱によって，素性によるクラスタリング結果の適合率と再現率が大きく異なる．

3.3 高精度クラスターに基づく半教師有り学習

我々は半教師有り学習の枠組みを用いて，関連度の強い素性から生成したクラスターに対して，関連度の弱い素性を用いて再現率を向上させる手法を提案する．

提案手法は，参照曖昧性解消の問題において一般的に適用することが可能な手法であり，関連度の強い素性とは参照の対象である実体に強い関連性を持つ素性である．提案手法は次の 2 段階からなる．

(1) 関連度の強い素性を元に文書間の類似度を計算し，クラスターを作成する．
(2) 関連度の弱い素性を元に文書間の類似度を計算し，クラスターに含まれている文書と高い類似性を持つ文書を収集し，クラスターに併合する．

本稿においては，参照曖昧性解消の問題として，Web 上の人名検索における同姓同名人物間の曖昧性を対象にする．第一段階のクラスタリングにおいて，クラスターを形成した文書とそれ以外の文書が存在する．第二段階のクラスタリングを半教師有り学習と考える時，前者が labeled data にあたり，後者が unlabeled data にあたる．

4. ブートストラップによる半教師有り学習

クラスターデータに対する半教師有り学習として実際に用いるブートストラップの手法について説明する．ここでは，人物との関連度が高い素性を用いて生成した，初期のクラスター集合において 2 つ以上の文書を含むクラスターを元にラベルを作成し，そこに含まれる文書を labeled data とする．そして，クラスターに分類されなかった文書を unlabeled data として，labeled data を元にラベルを割り当て，クラスターを拡張する．

4.1 アルゴリズム

ブートストラップのアルゴリズムは Algorithm 1 に示した．このアルゴリズムは共起行列 P を元に文書のクラスターへの帰属度行列 $\mathrm{R}_D^{(t)} = \{r_{d,C}\}$，素性のクラスターへの帰属度行列 $\mathrm{R}_F^{(t)} = \{r_{f,C}\}$ を反復計算し，クラスターの拡張を行う．以下にその詳細を説明する．ブートストラップのアルゴリズムは次の段階からなる．

(1) 2 行目: 文書と素性の共起行列 P を計算する．ここでは，文書と素性の共起行列が極めて疎であるため，自己相互情報量が 0 以下となるものは全て 0 としている．
(2) 4 行目: 共起行列 P と文書のクラスターへの帰属度行列 $\mathrm{R}_D^{(t)}$ を元に素性のクラスターへの帰属度行列 $\mathrm{R}_F^{(t)}$ を計算する．
(3) 5 行目: 共起行列 P と素性のクラスターへの帰属度行列 $\mathrm{R}_F^{(t)}$ を元に文書のクラスターへの帰属度行列 $\mathrm{R}_D^{(t+1)}$ を計算する．
(4) 8 行目: 各文書 d について帰属度 $r_{d,C'}$ が最大となるクラスター $C' \in \mathcal{C}'$ を選択する．$\mathcal{C}'$ は $\mathcal{C}$ に含まれるクラスターのうち，文書を 1 つ以上含む集合である．ただし，$\mathcal{C}'$ は初期クラスター集合 $\mathcal{C}^{(0)}$ において，d が属していたクラスターを含む．

このうち，(2)，(3) を繰り返し，得られた文書のクラスターへの帰属度を元にクラスター集合 $\mathcal{C}'$ を生成し，結果とする．

初期に与えられる文書集合 D，素性集合 F，文書のクラスターへの帰属度行列 $\mathrm{R}_D^{(0)}$ の要素は，初期のクラスター集合 $\mathcal{C}$ において，文書 d が集合 $C(\in \mathcal{C})$ に属している場合は $r_{d,C}^{(0)} = 1$ とし，それ以外の場合は $r_{d,C}^{(0)} = 0$ としている．

4.2 Espresso に基づく帰属度計算

本研究では，文書と素性間の共起行列 P の計算に，情報抽出におけるブートストラップ手法である Espresso[18] において用いられている自己相互情報量を用いる．

Vol.2009-DBS-148 No.6
Vol.2009-FI-95 No.6
2009/7/28

Algorithm 1 ブートストラップに基づく二段階クラスタリング

Procedure: $D, F, \mathrm{R}_D^{(0)}$
// D:文書集合, F:素性集合, $\mathrm{R}_D^{(0)}$:文書のクラスターへの帰属度行列
// 共起行列 P の計算

$$\mathrm{P}[d,f] = \begin{cases} \frac{1}{\mathrm{maxpmi}} \log \frac{p(d,f)}{p(d)p(f)} & \text{if } \frac{p(d,f)}{p(d)p(f)} > 1 \\ 0 & \text{otherwise} \end{cases}, (d \in D, f \in F)$$

where $\mathrm{maxpmi} = max(\mathrm{P}[d', f'])$ $(d' \in D, f' \in F)$
for $t \in 0, \cdots, T-1$ // T:ブートストラップの反復回数 **do**
  // 素性のクラスターへの帰属度行列の計算

$$\mathrm{R}_F^{(t)} = \frac{1}{|D|}\mathrm{PR}_D^{(t)}$$

  // 文書のクラスターへの帰属度行列の計算

$$\mathrm{R}_D^{(t+1)} = \frac{1}{|F|}\mathrm{P}^{\mathrm{T}}\mathrm{R}_F^{(t)}$$

end for
for $C \in \mathcal{C}$ **do**
  $C_d^{(T)} = \arg\max_C r_{d,C'}^{(T)}$ where $\{C' | (C' \in \mathcal{C} \wedge |C| > 1) \vee C_d^{(0)}\}$
end for
C_d を元に $\mathcal{C}^{(T)}$ を決定
return $\mathcal{C}^{(T)}$

Espresso を用いたブートストラップ手法はグラフに基づく行列計算として扱うことができ[13]，共起行列 P を元に文書の隣接行列 $\mathrm{A} = \mathrm{P}^{\mathrm{T}}\mathrm{P}$ を計算することで，文書のクラスターへの帰属度についての計算として，式 (3) のように扱うことができる．

$$\mathrm{R}_D^{(t+1)} = \frac{1}{|D||F|} \cdot \mathrm{AR}_D^{(t)} \tag{3}$$

5. 同姓同名問題における高精度クラスターの作成

この章では，ブートストラップを同姓同名問題に適用するための適合率の高いクラスターの作成手法について，特徴抽出，類似度計算，クラスタリングの観点から述べる．

5.1 特徴量抽出

5.1.1 固有表現抽出

文書から人物に関連した固有名詞である固有表現を抽出する．固有表現として，本研究では人名，地名，組織名を扱っている．

しかし，地名，組織名には特定人物との関連が弱く，複数の人物に共通する固有名詞が多く存在する．そのため，別のデータセット [⋆1]から計算した大域頻度に基づいて，大域頻度の高い固有名詞は取り除く．

5.1.2 重要語抽出

文書から検索対象となる人物に関連した単語・句を抽出する方法のもう 1 つである重要語を用いた抽出について説明する．

文書に対して，形態素解析を適用した結果から，Term Extract [⋆2] を用いて重要語を抽出する[17]．重要語抽出は以下のようにして行われる．まず，形態素解析の結果から名詞句 $\mathbf{w}$ を取り出し，$\mathbf{w} = \{w_1, w_2, \cdots, w_L\}$ に存在する各単語 w_i について，単語重要度 $\mathrm{LR}(w_i)$ を式 (4) に従って計算する．

$$\mathrm{LR}(w_i) = \sqrt{(\mathrm{LF}(w_i) + 1) \cdot (\mathrm{RF}(w_i) + 1)} \tag{4}$$

$\mathrm{LF}(w_i), \mathrm{RF}(w_i)$ は文書内の全ての名詞句 $\{\mathbf{w}\}$ 内において単語 w_i の前，後に単語が存在する回数であり，これに対して 1 を加えて平滑化を行う．

単語重要度 $\mathrm{LR}(w_i)$ を元にして，名詞句の重要度 $\mathrm{FLR}(\mathbf{w})$ を式 (5) に従って計算する．

$$\mathrm{FLR}(\mathbf{w}) = \mathrm{F}(\mathbf{w}) \cdot \left(\prod_{i=1}^{L} \mathrm{LR}(w_i) \right)^{\frac{1}{L}} \tag{5}$$

$\mathrm{F}(\mathbf{w})$ は文書中の名詞句 $\mathbf{w}$ の出現回数であり，L は名詞句の長さである．

このようにして，抽出した名詞句 $\mathbf{w}$ のうち，重要度 $\mathrm{FLR}(\mathbf{w})$ が閾値 θ_{CKW} 以上の名詞句を重要語とする．

5.1.3 リンク構造抽出

文書内に含まれる他文書へのリンクを抽出し，特徴量として用いる．文書の<a>タグに含まれる URL と文書自身の URL を抽出し，正規化を行った後，URL による特徴量とする．URL についても，固有表現と同様に大域頻度の高い URL を取り除く．

⋆1 実験では，WePS-1 の訓練データから作成．
⋆2 http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/resource/termext/atr.html

Vol.2009-DBS-148 No.6
Vol.2009-FI-95 No.6
2009/7/28

5.2 類似度計算

本研究では，階層併合クラスタリングを用いて，第一段階のクラスタを作成する．ここでは，階層併合クラスタリングに必要となる各文書間の類似度について説明する．

5.2.1 Overlap 係数の導入

各特徴量の類似度に用いる Overlap 係数[16] について説明する．Overlap 係数は式 (6) のように計算される．

$$\mathrm{Overlap}(d_x, d_y) = \frac{|\mathbf{f}_x \cap \mathbf{f}_y|}{\max(\min(|\mathbf{f}_x|, |\mathbf{f}_y|), \theta_{\mathrm{overlap}})} \tag{6}$$

$\mathbf{f}_x, \mathbf{f}_y$ はそれぞれ文書 d_x, d_y に含まれる特徴量の集合である．$|\mathbf{f}_x \cap \mathbf{f}_y|$ は文書 d_x, d_y の共通する特徴量の数であり，$\min(|\mathbf{f}_x|, |\mathbf{f}_y|)$ は文書 d_x, d_y の特徴量の数の最小値である．$\theta_{\mathrm{overlap}}$ は特徴量の極端に少ない文書の影響を減らすために定める分母の取りうる最小値であり，本研究においては $\theta_{\mathrm{overlap}} = 4$ とする．

5.2.2 各特徴量ごとの類似度計算方法

- 固有表現
  固有表現による類似度 $\mathrm{sim}_{\mathrm{NE}}$ は固有表現抽出を用いて抽出した人名（Person），地名（Location），組織名（Organization）を用いて式 (7) のようにして計算する．
  $$\mathrm{sim}_{\mathrm{NE}}(d_x, d_y) = \alpha_{\mathrm{P}} \mathrm{sim}_{\mathrm{P}}(d_x, d_y) + \alpha_{\mathrm{L}} \mathrm{sim}_{\mathrm{L}}(d_x, d_y) + \alpha_{\mathrm{O}} \mathrm{sim}_{\mathrm{O}}(d_x, d_y) \tag{7}$$
  式 (7) の $\mathrm{sim}_{\mathrm{P}}, \mathrm{sim}_{\mathrm{L}}, \mathrm{sim}_{\mathrm{O}}$ は各属性の Overlap 係数から計算する．$\alpha_{\mathrm{P}}, \alpha_{\mathrm{L}}, \alpha_{\mathrm{O}}$ は各属性（人名，地名，組織名）についての重みである（$\alpha_{\mathrm{P}} + \alpha_{\mathrm{L}} + \alpha_{\mathrm{O}} = 1$）．重みは $\alpha_{\mathrm{P}} \gg \alpha_{\mathrm{O}} > \alpha_{\mathrm{L}}$ として，訓練データを用いて定める．
- 重要語
  重要語による類似度 $\mathrm{sim}_{\mathrm{CKW}}$ は重要語抽出を用いて抽出した複合語を特徴量として，式 (8) のようにして計算する．
  $$\mathrm{sim}_{\mathrm{CKW}}(d_x, d_y) = \mathrm{Overlap}(d_x, d_y) \tag{8}$$
- リンク
  リンクによる類似度 $\mathrm{sim}_{\mathrm{URL}}$ は元の HTML ファイルに含まれる URL から式 (9) のようにして計算する．
  $$\mathrm{sim}_{\mathrm{URL}}(d_x, d_y) = \begin{cases} 1 \text{ if } & d_x, d_y \text{ 間に直接リンクがある} \\ \mathrm{Overlap}(d_x, d_y) & \text{それ以外の場合} \end{cases} \tag{9}$$
  文書 d_x, d_y 間に直接リンクがある場合は類似度を 1 とし，そうでない場合は Overlap 係数を用いて計算する．

5.2.3 複数の特徴量による類似度

各特徴量によって計算された類似度を元にして，文書の類似度を決定する．ここでは，各類似度の最大値を文書の類似度とする．類似度を計算する場合，元の類似度が同一の値域を持つことが必要になる．各特徴量の値域は $[0, 1]$ であり，必要条件を満たしている．

5.3 クラスタリング

クラスタリングには，クラスターを再帰的に併合していく階層併合クラスタリングを用いる．ここでは，群間平均法を用いて，クラスター間の類似度を比較し，クラスタリングを行う．

6. 評価実験

半教師有り学習の手法を用いた二段階クラスタリングの手法を英語の同姓同名の人物の文書集合に対して適用し，クラスタリングの結果を評価する．

実験の手法について説明する．まず，一段階目のクラスタリングの手法について説明する．一段階目のクラスタリングには第 5 章で作成したクラスターを初期クラスターとして用いる．前処理として，lxml [⋆1]，Automatic English Sentence Segmenter [⋆2] を用いて，HTML ファイルを 1 行 1 文形式のテキストファイルに変換する．次に，特徴抽出として，形態素解析・固有表現抽出・URL 抽出を行う．形態素解析には，Tree Tagger [⋆3]，固有表現抽出には，Stanford NER [⋆4] を用いた．また，形態素解析の結果を用いて，重要語抽出を行う．前処理から得た固有表現・重要語・URL を元に文書間の類似度を計算し，上記の手法を適用した．文書間の類似度は特徴量から計算した類似度の最大値とした．固有表現の類似度計算に用いる重みは WePS-1 データセットの訓練データを用いて，$\alpha_{\mathrm{P}} = 0.78, \alpha_{\mathrm{O}} = 0.16, \alpha_{\mathrm{L}} = 0.06$ とした．次に，ブートストラップを用いた二段階クラスタリングについて説明する．クラスタリングには上記の方法で作成したクラスターを初期クラスターとして用いる．素性には，文書に含まれる単語の 1-gram，2-gram のうち stopwords を取り除いたものを用い，文書中の出現頻度を元にして，共起行列 P を計算した．1-gram については出現頻度の計算に TF-IDF を用いた．IDF の計算には，Web 1T 5-gram [⋆5] を用いた．帰属度の計算の試行

⋆1 http://codespeak.net/lxml/
⋆2 http://www.answerbus.com/sentence/
⋆3 http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
⋆4 http://nlp.stanford.edu/software/CRF-NER.shtml
⋆5 http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13

Vol.2009-DBS-148 No.6
Vol.2009-FI-95 No.6
2009/7/28

回数として，1-gram については $T = 1, 2, 3$ での結果を示し，2-gram については $T = 1$ についての結果を示した．

評価のためのデータセットには，英語における同姓同名問題解消タスクである WePS の第 1 回目，第 2 回目のデータセット WePS-1 [*1]，WePS-2 [*2]を用いた．各データは検索エンジンにおいて，人名での検索結果の上位ページを取ってきたものであり，取得不可能なものも合わせて， WePS-1 は最大 100 ページ，WePS-2 は最大 150 ページである．人名の数はともに 30 である．データセットには人手で作成した同一人物のクラスタの正解データが存在する．これらのデータは 1 つの文書が複数の同姓同名の人物について述べている場合を許容しており，複数のクラスタに属する文書が存在している．

6.1 評価方法

クラスターの評価方法としては， Purity/Inverse Purity と extended B-Cubed 指標を用いた．どちらの指標についても F-measure により，総合的なシステムの性能を評価する．これらの評価方法は同一文書が複数のクラスタに属することを許容した場合の評価方法である．

Purity, Inverse Purity による評価方法は以下の通りである[4]．結果のクラスタ集合を $\mathcal{C} = \{C_1, \cdots, C_i, \cdots, C_N\}$, 正解のクラスタ集合を $\mathcal{L} = \{L_1, \cdots, L_j, \cdots, L_M\}$ とする．任意の 2 クラスタ C_i, L_j の精度 Precision(C_i, L_j) を，

$$\text{Precision}(C_i, L_j) = \frac{|C_i \cap L_j|}{|C_i|} \tag{10}$$

と定義する．このとき， Purity 及び Inverse Purity は，

$$\text{P} = \sum_i \frac{|C_i|}{N} \max_j \text{Precision}(C_i, L_j) \tag{11}$$

$$\text{IP} = \sum_i \frac{|L_i|}{N} \max_j \text{Precision}(L_j, C_i) \tag{12}$$

となり，Purity/Inverse Purity $\text{F}_{\text{P-IP}}$ は，

$$\text{F}_{\text{P-IP}} = \frac{1}{\frac{1}{2}\left(\frac{1}{\text{P}} + \frac{1}{\text{IP}}\right)} \tag{13}$$

と計算される．

Amigó ら[3] は Purity/Inverse Purity が同姓同名人物クラスタリングの評価指標として不十分であることを示し，extend B-Cubed 指標を提案した[3]．extended B-Cubed 指標について説明する．文書 e が属するクラスタリング結果のクラスタ，正解クラスタをそれぞれ $C(e), L(e)$ とする．

extended B-Cubed 指標を算出する際に用いられる文書 e, e' 間の Multiplicity Precision （MP），Multiplicity Recall （MR）は式 (14)，(15) に従って計算される．

$$\text{MP}(e, e') = \frac{\min(|C(e) \cap C(e')|, |L(e) \cap L(e')|)}{|C(e) \cap C(e')|} \tag{14}$$

$$\text{MR}(e, e') = \frac{\min(|C(e) \cap C(e')|, |L(e) \cap L(e')|)}{|L(e) \cap L(e')|} \tag{15}$$

これらの指標を用いて，extended B-Cubed Precision （BEP），extended B-Cubed Recall （BER）を式 (16)，(17) のように $\text{MP}(e, e'), \text{MR}(e, e')$ の平均値を取ることで求められる．

$$\text{BEP} = \underset{e}{\text{Avg}} \left[\underset{e'.C(e) \cap C(e') \neq \emptyset}{\text{Avg}} [\text{MP}(e, e')] \right] \tag{16}$$

$$\text{BER} = \underset{e}{\text{Avg}} \left[\underset{e'.L(e) \cap L(e') \neq \emptyset}{\text{Avg}} [\text{MR}(e, e')] \right] \tag{17}$$

extended B-Cubed F-measure は extended B-Cubed Precision, extended B-Cubed Recall を元にして，

$$\text{F}_{\text{BEP-BER}} = \frac{1}{\frac{1}{2}\left(\frac{1}{\text{BEP}} + \frac{1}{\text{BER}}\right)} \tag{18}$$

と求められる．

Purity/Inverse purity は WePS-1 で用いられた指標であり， extended B-Cubed は WePS-2 で用いられた指標である．本研究の実験の評価では，両方の指標による結果を表記する．

6.2 実験:提案方法によるクラスタリング

提案手法によるクラスタリング手法を比較した．WePS-1 データセットに関する結果を表 1 に，WePS-2 データセットに関する結果を表 2 に示す．

実験におけるベースラインとして， ALL IN ONE， ONE IN ONE， COMBINED を用いた．ALL IN ONE は全ての文書を 1 クラスタにする場合， ONE IN ONE は各文書を 1 文書 1 クラスタに分けた場合， COMBINED は ALL IN ONE と ONE IN ONE のクラスタを合わせた場合である． COMBINED において，各文書は ALL IN ONE と ONE IN ONE の 2 クラスタに属することになる．

ORIGINAL は第 5 節の手法を用いて作成した階層併合クラスタリングの結果を示している．比較対象として，この ORIGINAL に対して，我々の先行研究で用いた重要語による二段階のソフトクラスタリング[12] を適用した結果を QE に示した．BootStrap は提案手法を

*1 http://nlp.uned.es/weps/weps-1-data/
*2 http://nlp.uned.es/weps/weps-2-data/

Vol.2009-DBS-148 No.6
Vol.2009-FI-95 No.6
2009/7/28

表 1 WePS-1 データセットによる評価実験

Topic	BEP	BER	$F_{BEP-BER}$	P	IP	F_{P-IP}
Baseline						
ALL IN ONE	0.18	0.98	0.25	0.29	1.00	0.40
ONE IN ONE	1.00	0.43	0.57	1.00	0.47	0.61
COMBINED	0.17	0.99	0.24	0.64	1.00	0.78
First-Stage Clustering						
ORIGINAL	0.84	0.73	0.77	0.82	0.73	0.76
Second-Stage Clustering						
QE(Soft)	0.82	0.76	**0.77**	0.84	0.73	0.77
BootStrap, 1-gram, $T = 1$	0.82	0.76	0.77	0.83	0.72	0.76
BootStrap, 1-gram, $T = 2$	0.44	0.86	0.54	0.46	0.91	0.58
BootStrap, 1-gram, $T = 3$	0.27	0.91	0.38	0.33	0.95	0.48
BootStrap, 2-gram, $T = 1$	0.84	0.73	0.77	0.82	0.73	0.76
WePS top 5						
1st	0.67	0.81	0.71	0.72	0.88	**0.79**
2nd	0.68	0.73	0.68	0.75	0.80	0.77
3rd	0.68	0.71	0.67	0.73	0.82	0.77

表 2 WePS-2 データセットによる評価実験

Topic	BEP	BER	$F_{BEP-BER}$	P	IP	F_{P-IP}
Baseline						
ALL IN ONE	0.43	1.00	0.53	0.56	1.00	0.67
ONE IN ONE	1.00	0.24	0.34	1.00	0.24	0.34
COMBINED	0.43	1.00	0.52	0.78	1.00	0.87
First-Stage Clustering						
ORIGINAL	0.92	0.70	0.78	0.94	0.79	0.86
Second-Stage Clustering						
QE(Soft)	0.87	0.77	0.81	0.91	0.84	0.87
BootStrap, 1-gram, $T = 1$	0.89	0.82	**0.85**	0.93	0.87	0.89
BootStrap, 1-gram, $T = 2$	0.66	0.91	0.73	0.93	0.76	0.82
BootStrap, 1-gram, $T = 3$	0.53	0.95	0.63	0.65	0.96	0.74
BootStrap, 2-gram, $T = 1$	0.92	0.70	0.78	0.94	0.79	0.86
WePS top 3						
1st	0.87	0.79	0.82	0.91	0.86	0.88
2st	0.85	0.80	0.81	0.87	0.89	0.87
3st	0.93	0.73	0.81	0.95	0.81	0.87

用いて行った二段階クラスタリングの結果であり，帰属度の計算において用いた行列を表している．1-gram，2-gram は二段階のクラスタリングにおいて用いた素性を表しており，T は二段階クラスタリングの試行回数を表している．

一段階目の階層クラスタリングの結果は各データセットをテストセットとして，もう一方のデータセットを訓練セットとして，訓練セットにおいて評価値 F-measure(BEP-BER) が最大となる閾値を学習し，その閾値を用いてクラスタリングを行った結果を示している．

各データセットの結果について説明する．データセットには WePS-1[4], WePS-2[5] の上位チームの結果を併記した．

WePS-1 データセットの結果について説明する．提案手法による結果では 1-gram についての試行回数 $T = 1$ の結果が Recall を改善し，0.77(B-Cubed 指標) を示している．これは以前の手法を用いた結果である QE が示した 0.77 と同等の結果である．

WePS-2 データセットの結果について説明する．WePS-2 のデータセットは ALL IN ONE ベースラインが ONE IN ONE ベースラインに比べて良い性能を示していることより，平均的に大きなクラスタによって構成されていると考えられる．提案手法による結果では，WePS-1 データセットと同じように 1-gram の試行回数 $T = 1$ の結果が最高値 0.85(B-Cubed 指標) を示している．WePS-1 データセットと比較して結果が大きく改善されていることは QE の結果からも分かるように，データセットの性質によるものと推定される．

WePS-1 データセット，WePS-2 データセットの結果からブートストラップを用いることによって評価が改善できていることが確認できた．特に第一段階で Precision の高いクラスターを作成していた，WePS-2 データセットでは Recall を第二段階において大幅に改善することができた．

2-gram を素性として用いた場合，クラスタリング結果において改善が行われていなかった．このことは 2-gram が 1-gram に比べて疎な素性集合であることが原因と考えられる．また，1-gram を素性とした場合の結果において，試行回数を $T = 2, 3$ とした時，Recall の向上に対する Precision の低下の割合が大きくなり，F-measure の低下より性能が悪化したことが確認できる．この問題は人物との関連性の弱い素性が過大評価され，人物との関連性が弱い文書をクラスターに帰属させているためだと考えられる．Precision 低下に対する解決策として，Komachi ら[13] の論文で述べられている手法である，ノイマンカーネル，グラフラプラシアンを用いたアルゴリズムによる実験も行ったが改善は見られなかった．改善は見られなかった．

7. おわりに

我々は，人物との関連度が高い素性を用いることで作成した，高精度のクラスターに対して，半教師有り学習法であるブートストラップを用いて，クラスターの拡張を行い，適合率

Vol.2009-DBS-148 No.6
Vol.2009-FI-95 No.6
2009/7/28

の低下を押さえ，再現率を向上させようとした．クラスターの拡張には，情報抽出における代表的なブートストラップ手法として知られている Espresso を応用し，文書と素性の関係に適用した．

本研究で用いた手法の評価として，Web 上での人物検索に関するワークショップ WePS のデータセットを用いて実験を行い，評価を行った．その結果， WePS-1 データセットでは $F_{\text{BEP-BER}} = 0.77$ と従来手法と同等の結果となったが，WePS-2 データセットで $F_{\text{BEP-BER}} = 0.85$ を示し，既存手法を大きく改善する結果を示した．

今後の課題として，用いる素性の選択，複数クラスターの利用方法についての検討を行っていく．

参 考 文 献

1) Abney, S.: Bootstrapping, pp.360–367 (2002).
2) Abney, S.: Understanding the yarowsky algorithm, *Computational Linguistics*, Vol.30, No.3, pp.365–395 (2004).
3) Amigó, E., Gonzalo, J., Artiles, J. and Verdejo, F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints, *Information Retrieval*, pp. 1–26 (2008).
4) Artiles, J., Gonzalo, J. and Sekine, S.: The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task, *The SemEval-2007*, pp. 64–69 (2007).
5) Artiles, J., Gonzalo, J. and Sekine, S.: WePS 2 Evaluation Campaign: overview of the Web People Search Clustering Task. (2009).
6) Artiles, J., Sekine, S. and Gonzalo, J.: Web people search: results of the first evaluation and the plan for the second, *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pp.1071–1072 (2008).
7) Bagga, A. and Baldwin, B.: Entity-based cross-document coreferencing using the Vector Space Model, *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pp.79–85 (1998).
8) Balog, K., Azzopardi, L.A. and de Rijke, M.: UVA: Language Modeling Techniques for Web People Search, *The SemEval-2007*, pp.468–471 (2007).
9) Bekkerman, R. and McCallum, A.: Disambiguating Web appearances of people in a social network, *Proceedings of the 14th international conference on World Wide Web*, pp.463–470 (2005).
10) Chen, Y. and Martin, J.: CU-COMSEM: Exploring Rich Features for Unsupervised Web Personal Name Disambiguation, *The SemEval-2007*, pp.125–128 (2007).
11) Elmacioglu, E., Tan, Y., Yan, S., Kan, M. and Lee, D.: PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features, *The SemEval-2007*, pp. 268–271 (2007).
12) Ikeda, M., Ono, S., Sato, I., Yoshida, M. and Nakagawa, H.: Person Name Disambiguation on the Web by TwoStage Clustering, *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference* (2009).
13) Komachi, M., Kudo, T., Shimbo, M. and Matsumoto, Y.: Graph-based Analysis of Semantic Drift in Espresso-like Bootstrapping Algorithms, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp.1010–1019 (2008).
14) Liu, X., Gong, Y., Xu, W. and Zhu, S.: Document clustering with cluster refinement and model selection capabilities, *In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp.191–198 (2002).
15) Mann, G.S. and Yarowsky, D.: Unsupervised personal name disambiguation, *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pp.33–40 (2003).
16) Manning, C. and Schütze, H.: *Foundations of statistical natural language processing*, MIT Press (1999).
17) Nakagawa, H. and Mori, T.: Automatic term recognition, *Terminology*, Vol.9, No.2, pp.201–219 (2003).
18) Pantel, P. and Pennacchiotti, M.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations, *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pp.113–120 (2006).
19) Popescu, O.: IRST-BP: Web People Search Using Name Entities, *The SemEval-2007*, pp.195–198 (2007).
20) Saggion, H.: SHEF: Semantic Tagging and Summarization Techniques Applied to Cross-document Coreference, *The SemEval-2007*, pp.292–295 (2007).
21) Slonim, N. and Tishby, N.: Document clustering using word clusters via the information bottleneck method, *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp.208–215 (2000).
22) Tishby, N., Pereira, F.C. and Bialek, W.: The information bottleneck method, *Proceedings of the 37-thAnnual Allerton Conference on Communication* (2000).
23) Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods, pp.189–196 (1995).