社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2006－ＮＬ－174（１）
2006／7／27

組織情報を用いた人名の曖昧性解消方式

相薗　敏子
（株）日立製作所 中央研究所 〒185-8601 東京都国分寺市東恋ヶ窪 1-280
E-mail: toshiko.aizono.jn@hitachi.com

あらまし　人名文字列とそれに対応する実体としての「人」には同姓同名による曖昧性がある。本研究ではまず，テキストに出現する人名の曖昧性ついて営業日報データ 7,600 件を対象に調査を行った。その結果，営業日報データには延べ 5,778 件の人名が出現しており，そのうち 55%に同姓同名による曖昧性が存在し，文字列だけで「人」を同定すると最大 52 人の「人」を同一人物としてしまう可能性があることが分かった。これに対して，本研究では人名と同じ文に出現する組織名を利用した曖昧性解消アルゴリズムを提案する。先の営業日報データを用いた実験では，曖昧性のある人名に対して 89%の精度で正しく「人」に同定できるという結果を得た。
キーワード　人名，同姓同名，曖昧性，組織名，組織情報

A Method of Person Name Identification using Organization Information

Toshiko Aizono
Central Research Laboratory, Hitachi, Ltd. 1-280 Higashi-Koigakubo, Kokubunji-shi, Tokyo, 185-8601 Japan
E-mail: toshiko.aizono.jn@hitachi.com

Abstract　In this paper, I describe the identification issue of person name which appeared in text. I explore 5,778 person names which are extracted from 7,600 sales reports, 55% of them are ambiguous due to multiple candidates in identifiable person list. Also this result shows 52 people with the same surname at the maximum may be treated as one person. In order to resolve this problem, I propose an algorithm using organization name which co-occur with person name in the same sentence. In an experiment using the sales reports, 89% of the ambiguous person names are identified correctly.
Keyword　Person Name, Identification, Ambiguity, Organization,

1. はじめに

ITの普及により企業で作成／保存されている電子化情報の量は増加の一途にある。これら膨大な情報に対して企業では，適切に管理／運用すると同時に，知的生産性の向上などのために有効に活用したいというニーズがある。

このようなニーズに応えるには，保有する情報に対してメタデータを付与し，企業の業務知識などを定義したオントロジを用いてメタデータを解釈しながら情報を処理できる仕組みが有用であると考える。ここで情報に付与するメタデータには，作成日のようなコンテンツに依存しないものもあるが，情報をきめ細かく管理・活用するにはコンテンツを表すメタデータが付与されていることが望ましい。

一方で企業内の電子化情報の 80%は，報告書やメールのようなテキスト形式の情報であるといわれている。テキスト形式の情報には，コンテンツを表すメタデータとしてテキストに含まれる普通名詞や人名，地名のような固有表現が挙げられる。これらのうち本研究では，人名に着目した。その理由は，企業におけるテキスト情報には組織の活動が記録されており，コンテンツに含まれる人名は活動の主体や対象（「誰が」，「誰と」，「誰に」）を表している。また企業内には，一般に組織の構成や所属するメンバを記述した組織情報がある。このような組織情報をオントロジとして利用しテキストに出現する人名を組織情報に対応付けることで，組織における「人」の活動に関連付けて情報を管理／活用することが可能となると考えるからである。

一方で，テキストに出現する人名には同姓同名が存在するため，組織情報に登録された「人」との対応付けに曖昧性が生じる。特に企業内のテキストでは人名が名字のみで出現することが多く，「人」への対応付けに高い曖昧性が生じうる。このような同姓同名による曖昧性は，テキストに出現する人名を利用したシステムで精度を低下させる大きな要因となっている。一例として，人名をキーとした文書分類システム[1]，Web空間などからのコミュニティ情報やパーソナル情報の抽出[2][3]などでは，いずれも精度低下の要因のひとつとして同姓同名の存在が挙げられている。

本報告では，企業内のテキストに出現する人名について実データを用いて調査を行い，人名を組織情報に登録された「人」に同定するための曖昧性解消方式を提案する。

以下，第２章では企業内のテキストの一例として営業日報データを対象に出現する人名を調査した結果を示す。第３章では組織名を利用して人名の曖昧性を解消し「人」に同定するアルゴリズムを提案し，第４章で営業日報データを用いてアルゴリズムを実験した結果を示す。最後に第５章で関連研究との比較を述べる。

表1 営業日報に出現する人名文字列と人名 ID の数

項目	延べ数(件)	異なり数(件)	平均出現頻度(回)
人名文字列	5,778	1,426	1.88
人名 ID		2,622	2.20

表2 営業日報に出現する人名数(詳細)

#	項目	平均	最大	最小	例
1	文字列あたりの異なりID数	4.05	48	1	文字列「鈴木」に対応するID:「1001」,「1002」...(48件)
2	IDあたりの異なり文字列数	1.02	3	1	ID0313に対応する文字列:「鈴木」,「鈴木太郎」(2件)

2. 調査

2.1. 概要

まず営業日報データ7,600件(延べ単語数60万語)から人名を人手で抽出した。その結果,延べで5,778件の人名を得ることができた。次にそれらが出現する文脈を手がかりに文字列に対応する「人」をチェックし,各「人」に対して一意に職別可能なIDを付与した。その結果,異なりで2,622件の人名IDを得ることができた。

表1は,営業日報データに出現する人名と人名 ID の数を示したものである。表1に示すように営業日報データには,延べ5,778件の人名に対して異なり文字列数が1,426件であった。これは,文字列が同じならば同一人物であると見なすと,データには 1,426 人の「人」が出現したということを意味する。一方で,同じ人名でも別の「人」を指す可能性がある。表中,人名IDの異なり数2,622件は実際の「人」の数に相当する。その数は先の異なり文字列数の約 1.8 倍であった。

2.2. 考察

表1の結果は,単純に「文字列が同じなら同じ人を指す」として人名を扱った場合(1,426 件)と実際の「人」の数(2,622 件)との間には大きな誤差があることを表している。ここで前記誤差について詳細に考察するため,ひとつの文字列に対応する ID の数,およびひとつの ID に対応する文字列の数を求めた。その結果を表2に示す。

表2中,#1 の文字列あたりの異なり ID 数は,同じ人名に対して対応する「人」の数が平均4件あることを表す。最大は 48 件の対応先がある人名「鈴木」であった。これは「鈴木」という名字の「人」が 48 人出現していることを指す。なお同じ人名とは,文字列が完全一致しているものを指す。人名「鈴木」に「鈴木太郎」などフルネームで表記されている文字列も含めて異なりIDの数を求めると,計52件,すなわち52人の「鈴木」さんの出現を確認することができた。

また文字列あたりの異なりID数が1でない人名,すなわち対応先が複数ある人名は,延べで3,164件(総文字列数の55%)であった。これは,テキストから人名文字列を抽出し組織情報に対応付けようとすると,文字列の半数以上に曖昧性があることを意味している。

一方,表 2 中#2 の ID あたりの異なり文字列数は,同じ「人」に対して複数の人名文字列があることを指し,「表記ゆれ」に相当する。最大は3件であり,一例として「渡辺」さんは,前後の文脈から同一人物であることが推定できるのにもかかわらず,「渡邊」と「渡邉」の表記でも出現していた。なお「鈴木」と「鈴木太郎」のように,名字のみに加えてフルネームでも出現している場合も表記ゆれに含まれている。

ここで ID あたりの異なり文字列数が1でないもの,すなわち表記ゆれのある「人」の数を調べたところ60件あり,それに相当する人名文字列の数は延べで 592 件であった。これは,テキストから人名を抽出し文字列ベースで扱うと,同一人物であるにもかかわらず別人とされてしまうものが,全体の 10%を占めることを意味している。ただしその割合は,前記の対応先に曖昧性があるもの(55%)よりも低いことから,本報告では人名の表記ゆれについては検討の対象外としている。

3. 方式の提案

3.1. アプローチ

テキストに出現する人名文字列に関して次のような仮説が考えられる。以下,これら仮説について説明する。

仮説 1)人名文字列 Pi が出現する文には,Pi が所属する組織名が Pi の近くに出現する可能性が高い。

仮説 2)Pi と同じ人名文字列 Pj が類似文脈に出現していれば Pi と Pj は同一人物である可能性が高い。

(1)人名と同一文中に出現する組織名

企業内のテキストに含まれる人名文字列は名字のみで出現することが多く,かつ所属する部署や職名で修飾されることが多い。例を用いて説明する。

図1は,企業において人名を含むテキストと企業の組織情報の一例である。図中,テキストA,テキストB,およびテキスト C には,それぞれ人名「小島」が出現しており,一方で「国分寺ソリューション」という会社には「小島」という「人」が複数存在する。そのため人名「小島」は,対応する「人」に曖昧性がある。そのうちテキスト A には,図中①-1 に示すように人名「小島」の前方に組織名「営業1課」が出現している。またテキスト C には,①-2 に示すように人名の後方に組織名「営業3課」の略称「営3」が出現している。これらを手がかりとして図中②-1 および②-2 に示すように,「営業1課」で修飾された「小島」は営業1課の,「営3」で修飾された「小島」は営業3課の「小島」さんにそれぞれ対応付けることが可能である。

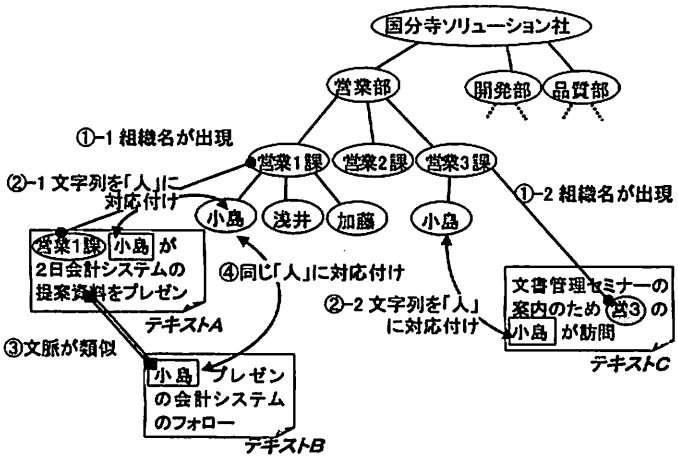

図1 テキストに出現する人名文字列の組織情報への対応付け

人名ID	人名文字列	所属先リスト
5201	鈴木	"開発部"
5202	鈴木	"営業部","営業1課"
5203	鈴木太郎	
…	…	…

図2 組織情報の一例

(2)類似した文脈に出現する人名

図1中，人名文字列と同一文中に組織名が出現するテキストAとテキストCに対して，テキストBには組織名が出現しないので，(1)の方法では曖昧性が解消できない。ここでテキストBの人名が出現している文脈を見ると，「会計システム」に関係していることが分かる。一方でテキストAとテキストCの文脈を見ると，テキストAの「小島」は「会計システム」に，テキストCの「小島」は「文書管理セミナー」に関係があり，テキストBはテキストAにより文脈が類似しているといえる(図中③)。これを手がかりとして，テキストBの「小島」はテキストAの「小島」と同一人物である可能性が高いと判定し，テキストBの人名文字列「小島」をテキストAの「小島」と同じ営業1課の小島さんに対応付ける(図中④)。

3.2. アルゴリズム

前節で述べた仮説に基づき人名文字列 Pi の人名ID Ii を判定する次の2つのアルゴリズムを提案する。

(1)アルゴリズムI

文字列から可能な人名ID Iiのリスト Isiを取得でき，かつ人名ID Ii から所属先のリスト Bi={bi1, bi2,,,,bin}を取得できる組織情報があるとき，同一文に出現する組織名を手がかりに人名 Pi の人名ID Ii を判定するアルゴリズム

初期状態:テキストの文字列位置を示すインデックス x に 0 (初期値)をセットする。

ステップ1:x を1増やして，x 番目から始まる文字列をキーに組織情報を最長一致検索する。一致する人名 Pi があれば当該人名の人名ID のリスト Isi={Ii1, Ii2 ,,,, Iim}を取得してステップ2へ。一致する人名がなければ，ステップ1の処理を繰り返す。

ステップ2:一致した人名 Pi がフルネームであり，かつ対応する人名IDリストIsi の要素数が1であれば，人名 Pi を人名IDリスト Isi の唯一の要素 Ii1 に対応付けてステップ1の処理に戻る。(フルネーム優先)

ステップ 3:Pi と同一文に出現する組織名のリスト Tsi={ti1, ti2, ,,, tik}を作成する。ここで各 tijにはPiからの距離に基づく重み(Pi に近いほど大きい)をつける。

ステップ 4:類似度を格納する変数 simmax に 0 をセットし，人名IDのリスト Isi の要素 Iij に対して，それぞれ以下の処理を行う

ステップ 4-1:組織情報から人名ID Iij の所属先リスト Bjk={bk1, bk2, ,,, bkl}を取得する。

ステップ 4-2:所属先リスト Bjk と組織名リスト Tsi との類似度 sim(ベクトルのコサイン値など)を求める。

ステップ 4-3:類似度 sim が 閾値 α より大きく，かつ simmax より大きければ，sim の値を simmax にセットしてIij の値を q にセットする。

ステップ 5:simmax が0よりも大きいなら，q にセットされた人名 IDIij を Pi の人名IDとし，ステップ1の処理に戻る。

上記において，Pi と同一文中に出現する組織名がなければ(Tsi= φ)，Pi はどの人IDにも対応付けられない。また可能な人名IDの数が1つしかなくても，Tsi との類似度 sim が閾値 α 以下であれば対応付けない。

なお，上記の組織情報において，各人名ID Ii の所属先リスト Bi および Pi と同一文中に出現する組織名リスト Tsi には，組織を構成する部署名以外に，職位名(「課長」など)，あるいは担当業務名(例:「SE」など)等を含んでもよい。ただしその場合，職位／業務名には部署名と比して小さな重みをつけるようにし，ステップ 4-2 において組織名リスト Tsi との類似度 sim を求める際，職位名／業務名しか一致していない場合には，sim は0とするなどの工夫をすべきである。

(2)アルゴリズムII

テキストの集合Dがあり，その各要素に対してアルゴリズムIを適用した結果，人名IDが判定できない人名があるとき，出現文脈の類似性に基づいて Pi の人名ID li を判定するアルゴリズム

ステップ1:Pi の出現文脈 Tdi={ti1,ti2,,,,tik}を作成する。ここで tik は，Pi と同じ文書に出現する名詞類(普通名詞，動作名詞，固有名詞，英字列，未知語など)とする。また，tik には出現頻度あるいは出現頻度に出現文書数の逆数をかけた値(tf/idf値)などを重みとしてつけてもよい。

ステップ2:テキストの集合D中，Piと同じ文字列の人名を取得して人名リスト Psi={P1,P2,,,,Pn}を作成する。

ステップ3:類似度を格納する変数 simmax に0をセットし，人名リストPsiの要素 Pj(j=1,2,,,,n)に対して以下の処理を行う。

ステップ3-1:Pjの出現文脈 Tdj={tj1,tj2,,,,tjm}を Piと同様に作成する。

ステップ3-2:Tdi と Tdj との類似度 sim(ベクトルのコサイン値など)を求める。

ステップ3-3:類似度 sim が閾値 β より大きく，かつ simmax より大きければ，sim の値を simmax にセットして Pj の値を q にセットする。

ステップ4:simamx が0よりも大きいなら，q にセットされた人名 Pjと Piを同一人物とする。

ステップ5:Pj に人名ID Ij が付与されていれば当該人名ID Ij を Pi にも付与する。

上記のうち，テキストの集合D中，Pi と同じ文字列の人名がない場合，あるいはその出現文脈と Pi の出現文脈の類似度が β よりも小さければ，Pi は組織情報には登録されていない人名であるとする。

4. 実験

4.1. 概要

第2章の調査で抽出した人名文字列のうち対応先に曖昧性があるものを対象に前章で示したアルゴリズムIおよびIIを適用し，「人」が登録された組織情報に人名を対応付ける実験を行う。以下，本実験で用いる人名と組織情報，および実験課題について説明する。

(1)実験対象

第2章で示したように，営業日報データから抽出した人名文字列 5,778 件(文字列の延べ数。以下同じ)のうち対応する人名 ID に曖昧性のあるものは 3,164 件であった。このうち 833 件は，営業日報を作成した会社の社員以外の間での曖昧性，すなわち顧客などの間で同姓が存在することによる曖昧性であった。

本実験では，営業日報データに出現する人名文字列を後述する組織情報に登録されている人名 ID に対応付けることを課題とする。この実験課題において，前記 833 件は候補となる人名 ID が組織情報に含まれていないため，アルゴリズムの適用対象から外れる。そのためこの 833 件は実験では除き，残る 2,331 件を評価の対象に用いた。

(2)組織情報

本実験では，先の営業日報データを作成した会社のまとまった組織情報が入手困難であることから，実験者がデータを参照して組織情報を作成した。具体的には人名が出現した文脈において，当該人名の所属を表す単語のリストを当該人名の組織情報とした。一例を図2に示す。

図2中，人名 ID「5201」および「5202」は，人名「鈴木」に対応する ID であり，その組織情報は「5201」が「開発部」，「5202」が「営業部」および「営業1課」である。一方で，人名が出現する前後の文脈(例:「弊社鈴木がご案内した」)によって組織の一員であることは判定できるが，具体的な所属先が不明な人名が 821 件中 33 件あった。そのような場合，組織情報は，図2の人名 ID「5203」のように空集合とした。

(3)実験課題

実験対象の 2，331 件の人名を，作成した組織情報に登録されている人名 ID に対応付けることを課題とする。ここで 2,331 件の人名文字列のうち，組織内の人名は 821 件である。それ以外の 1,510 件は，組織内の人名と同姓の社外者(顧客など)の人名である。従ってこれら 1,510 件は対応する人名 ID なしと判定することが課題となる。

4.2. 結果および結果の検討

実験の結果を表3および表4に示す。表3は実験対象である2,331件の人名文字列に対してアルゴリズムIを適用した結果，表4はアルゴリズムIに加えてアルゴリズムIIを適用した結果である。

(1)アルゴリズムIの結果の検討

表3からアルゴリズムIのみでも正解率が 89.0%と，高い精度を得ることができた。このことから，人名と組織名は同じ文中近くに出現しやすいということが，少なくとも営業日報データでは確認することができた。

実験対象の 2,331 件の人名のうち，同一テキスト内に出現している人名の間に曖昧性があるケース，すなわち同姓の別人がひとつのテキストに出現しているケースが 10 組，人名数で 20 件あった。これらに対してもアルゴリズムIは，9組／18 件の人名に正しい ID を付与することができた。

また，4.1 の(2)で述べた所属が不明な人名 33 件のうち，26 件は正しい ID を付与することができた。これら 26 件は，フルネームで出現しており，かつ同姓同名が組織情報に存在しないため，同一文中に組織名が出現していないもののID を付与することができた。

一方で，アルゴリズムIの誤りのうち，ほとんどが ID の判定不可(正解では人名 ID があるにもかかわらず ID を判定できなかったもの，9.4%)であった。この誤りに関し人名が出現する文を調べたところ，その要因の多くは，人名が出現する文に適切な組織名が出現していないことにあった。

表3 アルゴリズムⅠの実験結果

項目		数(件)	割合(%)	備考
正解		2,075	89.0	
誤り	判定不可	219	9.4	正解の人名 ID があるにもかかわらず ID を付与できなかったもの
	ID 対象外	35	1.5	正解は人名 ID がない*にもかかわらず ID を付与したもの
	ID 不一致	2	0.1	付与した ID が正解の ID と異なる
合計		2,331	100.0	

*社外の人名など

表4 アルゴリズムⅠ＋アルゴリズムⅡの実験結果

項目		数(件)	割合(%)	備考
正解		2,087	89.5	
誤り	判定不可	184	7.9	正解の人名 ID があるにもかかわらず ID を付与できなかったもの
	ID 対象外	57	2.5	正解は人名 ID がない*にもかかわらず ID を付与したもの
	ID 不一致	3	0.1	付与した ID が正解の ID と異なる
合計		2,331	100.0	

*社外の人名など

表5 アルゴリズムⅠとアルゴリズムⅠ＋アルゴリズムⅡの比較

項目		数(件)	割合(%)	備考
結果が同じ	共に正解	2,053	88.1	
	共に誤り	222	9.5	共に「判定不可」など
	小計	2,275	97.6	
結果が異なる	正解(正解率向上)	34	1.5	アルゴリズムⅠでは ID 不明であったがアルゴリズムⅡで正しい ID を付与できたもの
	誤り(副作用)	22	0.9	アルゴリズムⅠで正解であったにもかかわらずアルゴリズムⅠで誤りとなったもの*
	小計	56	2.4	
合計		2,331	100.0	

＊社外者のため ID なしと判定すべき文字列に対して，組織のメンバの ID を付与したものが多い

アルゴリズムⅠは人名と同じ文に出現している組織名に基づいて ID を判定するため，前記の組織名の不在には対応することができない。そのため前記 ID の判定の不可はアルゴリズムⅠの限界を示しており，アルゴリズムⅠによる正解率の上限値はおよそ 90%前後であると推定できる。

ただし，本実験では組織情報に対応付ける際に曖昧性がある人名文字列のみを対象としている。営業日報データに出現する人名のうち対応先に曖昧性がないものを含めると，営業日報データに出現した人名のうち 95%に対して正しく処理できると予想する。

一方で本実験では，組織情報として人名と同一文中に出現する組織名のリストを利用している。そのため，人名 ID に対応する組織情報と，人名と同一文中に出現する組織名は一致する可能性が高く，その結果，高い精度が得られたといえる。この点を考慮し，今後は組織情報の入手／作成の方式や課題などについて検討する必要がある。

(2)アルゴリズムⅠ＋Ⅱの結果の検討

表4からアルゴリズムⅠとアルゴリズムⅡを適用した結果，正解率が 89.5%となった。この値は，表3に示すアルゴリズムⅠのみの正解率(89.0%)と大きな差はなく，本実験において出現文脈の類似性に基づくアルゴリズムⅡの効果は認めることができなかった。

表4の誤りを表3の誤りと比較してみると，ID 不明の誤りの割合が減少しているものの，一方で ID 対象外(正解の ID がないにもかかわらず ID を付与したもの)の割合が高くなっていることがわかる。

ここで，アルゴリズムⅠのみの結果とアルゴリズムⅠ＋Ⅱの結果を詳細に比較した。その結果を表5に示す。表5中，アルゴリズムⅠとアルゴリズムⅠ＋Ⅱで結果が異なる文字列は 56 件ある。これらのうち，34 件はアルゴリズムⅡによって正しい ID が付与されており，精度向上に貢献している。一方で 22 件は，アルゴリズムⅠで正解，すなわち社外の人名であるため ID を付与すべきでないにもかかわらず，社内の人名と類似した文脈に出現しているため社内の人名の ID を付与してしまった文字列であり，アルゴリズムⅡの副作用である。

すなわち，アルゴリズムⅡはアルゴリズムⅠで ID を付与できない人名文字列に対して1%程度正解率を向上させる効果はあるものの，同時に同程度の副作用も発生させているため，全体としてはアルゴリズムⅠとの差異を認めることができなかった。

5. 関連研究との比較

テキストから人名を抽出する研究は，従来から固有表現抽出の一部として数多く報告されているが [4][5][6]，それに比して人名の曖昧性の解消に関する研究は多くはない。

佐藤[7]らは，Web ページをクラスタリングすることにより Web ページに出現する人名文字列を識別する方式を提案している。具体的には，識別対象の人名文字列(ただし，フルネーム)をキーに Web ページを検索し，検索結果を URL

を用いてクラスタリングし，さらに各クラスタからクラスタの特徴語として識別対象以外の人名文字列（やはりフルネーム）を抽出して同じ人名文字列が出現するクラスタをまとめていく。最終的に得られたクラスタが複数あれば，各クラスタ内に含まれる識別対象の人名文字列は同一人物，クラスタが異なれば含まれる人名文字列は別人であると判定する方式である。実験では，20件の著名人の人名文字列に関して，それぞれ9割以上の高い率で同姓同名の「人」を識別できたと述べている。

佐藤らの方式はクラスタの類似性，すなわち人名文字列が出現した文脈の類似性を利用している点で本報告で述べたアルゴリズムⅡと類似している。しかしこれら2つの方式は，前者がWebページのURLも文脈として利用しているのに対して，後者はテキストのみを文脈としている点で異なる。また，前者は文脈の類似性のみに基づくのに対して，後者は文脈の類似性は組織名に基づくアルゴリズムを補完するという位置付けである点で異なる。

松平らの研究[8]では，テキスト内に出現する人の属性情報を利用して人を識別する方式を提案している。本報告で述べたアルゴリズムは人の属性情報として組織名を利用しており，松平の研究とアプローチが近い。ただし松平らの研究では方式の提案のみであったのに対して，本報告では実験によってそのようなアプローチが有効であることを確認することができた。

6. まとめと今後の課題

本研究では，テキストに出現する人名の同姓同名による曖昧性を解消することを目的として，人名文字列と同じ文に出現する組織名を利用したアルゴリズムを提案した。営業日報データに出現する人名のうち，対応先に曖昧性がある2,331件の人名文字列を対象に実験を行ったところ，89%の精度で対応先を判定することができ，提案方式の有効性を確認することができた。

一方で，今後の課題として3点がある。

(1)同一文中に組織名が出現しない人名への対応

本報告で示したアルゴリズムⅠは高い精度が得られる一方で，同一文中に組織名が出現しない人名には対応することができない。そのため本報告では，人名が出現する文脈の類似性に基づくアルゴリズムⅡも提案した。前述の組織名を手がかりとしたアルゴリズムⅠに加えて文脈の類似性に基づくアルゴリズムⅡも適用した結果，わずかながら精度が向上することを確認した。ただし副作用も多く発生し，本報告の範囲では前記文脈の類似性に基づくアルゴリズムⅡが有効であるとは結論できなかった。今後は，同一文中に組織名が出現しない人名に対する他のアプローチを検討する必要がある。

(2)組織情報の自動生成

本報告では，人手で作成した組織情報を実験に用いたが，そのコストは少なくはない。これに対して，ある程度，既存の情報を利用して組織情報を自動的に作成する仕組みが必要である。

(3)異表記への対応

第2章でも述べたが人名または組織名の文字列は，組織情報に登録されているものとは異なる表記，例えば略称や異字などでテキストに出現する可能性がある。これに対しては，登録されている表記から異表記を自動的に生成することで対応可能であり，その方式の詳細および効果について実験／評価する必要がある。

文　献

[1] 増田恵子，梅村恭司，“固有名詞に着目し記事群を整理分類し提供するシステム”，情報処理学会，自然言語処理研究会報告，1996-NL-114，Vol.1996，No.65，pp.7-12，Jul. 1996.

[2] 風間一洋，原田昌紀，佐藤進也，福田健介，川上浩司，片井修，“人名を用いたWeb空間のコミュニティの解析”，情報処理学会，知能と複雑系研究会報告，2004-ICS-136，Vol.2004，No.85，Aug.2004

[3] 森純一郎，松尾豊，石塚満，”語の共起情報を用いたWeb上からの個人メタデータ抽出”，人工知能学会，セマンティックウェブとオントロジー研究会資料，SIG-SWO-A403-01，Nov.2004.

[4] 関根聡，井佐原均，“IREX：情報検索、情報抽出コンテスト”，情報処理学会，報学基礎研究会報告，1998-FI-051，Vol.1998，No.81，Sep.1998.

[5] 久光徹，丹羽芳樹，“辞書と共起情報を用いた新聞記事からの人名獲得”，情報処理学会，自然言語処理研究会報告，1996-NL-118，Vol.1997，No.29，Mar. 1997.

[6] 松尾衛，宮本昌幸，森辰則，“機械学習と人手作成のパタンを組み合わせた固有表現抽出”，情報処理学会，報学基礎研究会報告，1999-FI-057，，Vol.2000，No.29，Mar.2000.

[7] 佐藤進也，風間一洋，福田健介，村上健一郎，“実世界指向Webマイニングによる同姓同名人物の分離”，情報処理学会論文誌，Vol. 46，No. SIG 8 (TOD26)，pp. 26-36，Jun. 2005..

[8] 松平正樹，上田俊夫，渕上正睦，大沼宏行，森田幸伯，“文書からのキーワード抽出と関連情報の収集”，人工知能学会，セマンティックウェブとオントロジー研究会資料，SIG-SWO-A303-02，Mar. 2004.