4J－4

電子メールからのパーソナル情報抽出方法の検討
—住所録作成支援への適用—

浅野 久子　大山 芳史
NTT情報通信研究所

1．はじめに

電子メールには、末尾にsignatureが付与される場合が多い。このsignatureには送信者の姓名や電話番号等のパーソナル情報が含まれている。また、メールのheaderは送信者のメールアドレスを必ず含み、さらに姓名等の情報を含む場合もある。しかしsignatureの表現形式は多彩であり、含まれる属性数も様々で、パーソナル情報の抽出は単純なパターンマッチでは不十分である。

我々はインターネット上の電子メールを対象に、header, signatureを自動検出し、そこからパーソナル情報を抽出して住所録の作成支援をする研究を行っている。このうち本稿では、特にパーソナル情報の抽出方法を中心に述べる。

パーソナル情報の抽出は、テキストから特定の内容を抽出する内容抽出処理の一種と考えられる。しかし従来の内容抽出処理では対象テキストは新聞記事などに限定されおり[1]、外枠などの飾り用の文字が使われ、抽出すべき内容（パーソナル情報）もデザイン的に配置される文字列となるsignatureには適用できない。また、抽出対象となる社名等は膨大な種類があり、かつ増え続けており、単語辞書にすべて登録することは不可能である。そこで姓名辞書のみをもち、特定の単語群をキーワードとして扱うことを特徴とするパーソナル情報抽出処理を提案する。

2．処理の構成と概要

本処理は以下の４つの処理部で構成される。

（１）制御部

対象電子メールの選択、確認／自動住所録登録モードの選択等の各種設定、および（２）～（４）の処理の実行をGUIを介して行う。

（２）header, signature検出部

電子メールからheader（To, From, Subject等の情報をもつ）とsignature（電子メール送信者の署名。通常、本文末尾に付与される。）を検出する。

headerはフォーマットが明確に定義されており、検出は容易である。しかしsignatureは、存在するとは限らない、表現形式が多彩で本文との境界を検出するのが難しい等の課題がある。本処理では行単位のレイアウト情報、句読点情報、空行情報を利用することにより、これらの課題を典型的なsignatureに対してはほぼ解決することができた。

（３）パーソナル情報抽出部

（２）で検出したheaderとsignatureから、姓名や電話番号等のパーソナル情報の属性値を抽出する。詳細は３．パーソナル情報抽出処理で述べる。

（４）住所録登録部

住所録（姓名、会社名、住所などのフィールドをもったスタック）の各フィールドへ（３）で抽出したパーソナル情報の各属性値を登録する。

登録は（３）で抽出した属性値をユーザーが確認しながら住所録に登録する確認登録モードと、確認せずに自動的に登録を行う自動登録モードがある。

3．パーソナル情報抽出処理

headerとsignatureの情報を利用して、パーソナル情報の各属性値の抽出を行う。本抽出処理では以下の属性を抽出対象とする。

［抽出対象とする属性］姓名、会社名、所属[1]、メールアドレス、郵便番号、住所、電話番号、FAX番号

ここで、パーソナル情報抽出を行う際の課題となるsignatureの特徴として以下の(a)～(e)がある。

(a)飾り用の文字が多用される。
(b)何種類の属性が含まれるかは個々のsignatureごとに異なる。
(c)各属性の現れる位置・順序が多彩である。
(d)抽出対象ではない情報が存在する場合がある（例：個人の好きな言葉、趣味、愛称等）。
(e)文字列がデザイン的に配置されている（例：文字間にスペース等が挿入される）。

これらの各課題を解決するための処理のポイントを(A)～(E)で述べる。

(A)signatureの飾り文字検出（(a)対応）

縦または横に連続する記号文字、および任意の同一文字を飾り文字として検出する。検出した文字は抽出対象から除く。

(B)signatureのスコープ設定（(e)対応）

スコープ区切り記号（改行、スペース、(A)で検出した飾り文字、・，－，’，（，），〒以外の記号文字）で区切られる文字列をスコープと定義し、スコープ単位に属性値抽出処理を行う。また、スペースで区切られた記号以外の１文字からなるスコープが同一行に２つ以上連続して存在した場合、この連続するスコープを統合して１スコープとする。

(C)属性値抽出

(C-1)属性値抽出順序（(b),(c)対応）

処理の効率を考え、表現の定型度が高い属性から

Extraction of personal information from e-mail
—an application to support of making address books—
Hisako Asano, Yoshifumi Ooyama
NTT Information and Communication Systems Labs.

[1]会社名より下位の組織名

抽出を行い、抽出した文字列は以降の抽出では処理対象から除く。以下に抽出順序を示す。

メールアドレス→電話,FAX番号→郵便番号,住所→姓名→会社名,所属→住所（ビル名等）

姓名を会社名より先に抽出するのは、姓名があり会社名がないsignatureはありえるがその逆はまずない（異なる属性としての抽出や処理のfeedbackを抑止）、姓名の方が多義を絞り込みやすいためである。

(C-2)headerからのメールアドレス、姓名情報抽出

メールアドレスはheaderのFrom行より抽出する。また、メールアドレスのアカウント名部分("@"前方文字列）とメールアドレス以外の文字列により、姓名情報を得る（ローマ字の場合には、ローマ字→カナ変換を行い姓名の読み情報を得る）。

（例）From: Taroh YAMADA <yamada@abc.ac.jp>
下線部より"タロウ","ヤマダ"という読みを得る。

(C-3)キーワードパターン(KP)の利用（(d)対応）

各属性に対して、KP（任意文字指定や繰り返し指定など正規表現に準じる表現が記述でき、半角と全角、大文字と小文字を区別しないパターン表現）を作成した。そしてKPを含むスコープ、あるいは近隣スコープをその属性値として抽出する（抽出すべきスコープの位置は各属性ごとに定めている）。KPの総数は５５である。

（例）電話番号KPの１例："TEL[A-Z]*"(perl風表現)
"telephone", "TEL", "Tel", "tel", "ＴＥＬ", "Ｔｅｌ"のすべてにマッチ

(C-4)姓名辞書、headerの姓名情報の利用（(c),(d)対応）

姓名辞書（約２０万語）により、signatureから姓名を抽出する。複数の姓名が抽出された場合には、headerより得られた姓名情報を利用する。

このパーソナル情報抽出処理の例を図１に示す。

4．パーソナル情報抽出処理の評価と考察

評価対象となるsignatureは、電子メールを利用している人５名が、無作為に抽出した名刺１０枚を元に作成した５０signature（どの属性をsignatureに入れるかは各人の判断による）と、評価用に収集したデザイン的に凝っていると思われる３４signatureの計８４signatureである。ここで、本評価方法ではheaderが存在しないため(C-2)の処理は行わず、signatureのみでメールアドレス、姓名情報の抽出を行った。

評価結果を表１に示す。誤り率の算出は以下の式で行った。ここで、出現数とはその属性を含むsignatureの総数を表す。

抽出誤り率=その属性を誤って抽出したsignature数／出現数
未抽出率=その属性を抽出できなかったsignature数／出現数

表１より表現の多様性がある会社名、所属、住所が単語辞書を用いずKPのみの利用でかなり抽出できたといえる。これは会社名+所属や住所は階層的な構造であり、１スコープでもKPがマッチすればその近隣のスコープも同属性として抽出可能なためである。

また、主な誤り原因は、その属性値がアルファベット表記のみで表されている（15件, 抽出誤りの19%, 未抽出の50%）、姓名辞書に登録されていないような特殊な姓名である（9件, 抽出誤りの29%,未抽出の14%）、所属の一部は抽出できたが最下層（例：電気工学専攻M1）が抽出できない（5件, 抽出誤りの24%）などであった。

以上より、アルファベット表記、姓名辞書にない姓名への対処が課題であるといえる。

表１　評価結果

	出現数	抽出誤り率[%]	未抽出率[%]
姓名	83	8.4 (7)	7.2 (6)
会社名	71	11.3 (8)	12.7 (9)
所属	43	14.0 (6)	16.3 (7)
ﾒｰﾙｱﾄﾞﾚｽ	83	0 (0)	0 (0)
郵便番号	9	0 (0)	0 (0)
住所	14	0 (0)	0 (0)
電話番号	31	0 (0)	0 (0)
FAX番号	24	0 (0)	0 (0)
total	358	5.9 (21)	6.1 (22)

※誤り率の()内の数字は誤り個数

5．おわりに

今後は住所録作成支援システム全体としての評価を行っていく。

参考文献

[1]松尾, 木本：抽出パターンの階層的照合に基づく日本語テキストからの内容抽出方法, 情報処理学会論文誌, Vol.36, No.8, 1995

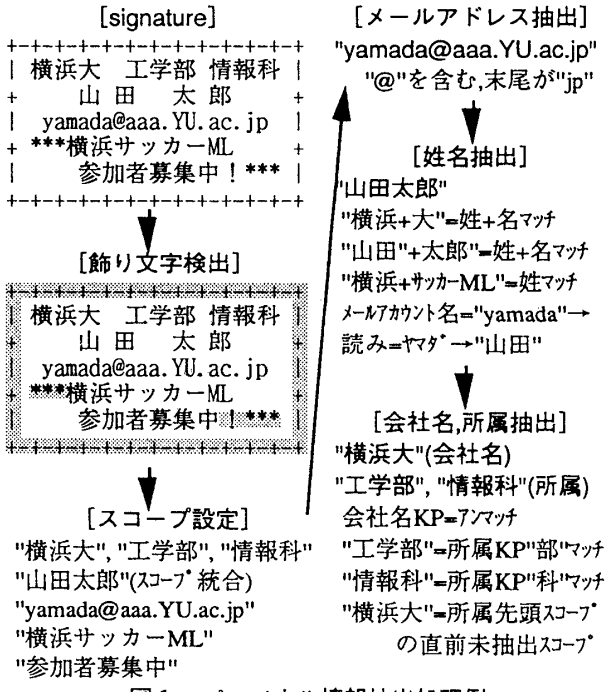

図１　パーソナル情報抽出処理例