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

2005－ＮＬ－168（10）
2005／7／22

プライバシ保護に向けた固有表現処理技術

小作　浩美[†]　相良かおる[††,†]　納谷　太[†]　桑原　彰教[†]　阿部　明典[†]　小暮　潔[†]

† ATR 知能ロボティクス研究所　〒619–0288 京都府「けいはんな学研都市」光台 2–2–2
†† 西南女学院大学　〒803–0835 北九州市小倉北区井堀 1–3–5
E-mail: †{romi,naya,kuwahara,ave,kogure}@atr.jp, ††sagara@seinan-jo.ac.jp

あらまし　本論文では，医療看護現場で収録した音声データの書き起こしテキストが与えられたとき，プライバシ保護を目的とした個人情報の抽出，および，データの分析用途に合わせて個人情報を自動的に抽出し変更するための手法について示す．まず，医療現場にて収集した看護師の会話データの特徴を示し，それに応じた固有表現抽出の方法について議論する．その結果，我々は，人名の表記パターンを利用した固有表現抽出技術を採用し，データから人名を抽出する実験を行った．結果として，精度 91.2%，再現率 94.7%が得られた．
キーワード　音声対話コーパス，プライバシ保護，固有表現抽出技術

Named Entity Recognition on Privacy Protection

Hiromi itoh OZAKU[†], Kaoru SAGARA[††,†], Futoshi NAYA[†], Noriaki KUWAHARA[†],
Akinori ABE[†], and Kiyoshi KOGURE[†]

† ATR, Intelligent Robotics and Communication Laboratories　2–2–2, Hikaridai, Keihannna Science City, Kyoto, 619–0288 Japan
†† Seinan Jo Gakuin University　Ihori 1–3–5, Kokura Kita-ku, Kitakyushu City, Fukuoka, 803–0835 Japan
E-mail: †{romi,naya,kuwahara,ave,kogure}@atr.jp, ††sagara@seinan-jo.ac.jp

Abstract　In this paper, we propose an automatic named entity recognition and extraction method. It is designed to recognize personal information for privacy protection, and according to situations to automatically extract and hide personal information. Especially we focus on the method that can be applied to medical applications. First, we show features of spoken dialogue data by nurses collected in a hospital, then we discuss a method to extract personal information from the data. We adopted a named entity recognition method by using a pattern dictionary for name expressions. We did name extraction experiments to the above data to obtain a recall rate 94.7%, a precision rate 91.2% in average.
Key words　Spoken dialogue corpus, Privacy protection, Named entity recognition

1. はじめに

医療において重要な問題として医療事故がある．例えば，米国においては医療事故による死亡者数が交通事故による死亡者数を上回ると推計されている [1]．そのため，医療事故を防止するための研究が行われている．医療事故を防止するためには，実際の医療活動がどのようになっているかを理解し，事故の原因を分析する必要がある．そのためには，医療看護現場での医療従事者などの対話を記録し，情報の流れを分析することが考えられる．我々はユビキタス・コンピューティング技術に基づく看護師支援技術の研究開発 [2] の一環として，医療看護現場での看護師の活動を様々なセンサーを用いて記録し，その分析を進めている．本稿では，音声データについて述べる．

従来の対話や会話データの研究においては，主に実験室環境における収録データを対象としていた．また，個人情報と関係しない話題を対象としたり，対話参加者の同意書を取ることにより，収集データ中の個人情報は問題とはなっていなかった．しかし，医療看護現場での音声データの収録では必然的に患者などの個人情報が含まれることになる．また，実際の現場で音声収録を行うと，収録対象者以外の人間の音声も収録されることがある．例えば，廊下での看護師同士の対話を収録する際，たまたま通りかかった患者を呼び止めたり，患者と会話をした音声が収録されるというようにである．従って，個人情報を不必要に漏洩せずに実験データとして使用するためには，収録

データから個人を特定する情報を削除することが求められる．例えば，対話音声の書き起こしにおいては，個人の特定に結びつく文字列を他の文字列に自動的に置き換えることが望まれる．そのためには，名前のような個人を特定するのに利用される情報（文字列）の自動抽出が必要になる．

人名や組織名を抽出する技術として，固有表現抽出技術が研究されている [3]．従来，固有表現抽出の技術は新聞記事からの人名や組織名の抽出を目的としており，話し言葉を対象とした研究は少なく，実際の現場で収録されているデータに対して，直接応用できるか定かではない．また，個人を特定する情報として，何を抽出するべきか，抽出すべき固有表現のタイプも明確にする必要がある．

本稿では，医療看護現場で収録した音声データの書き起こしテキストが与えられたとき，プライバシ保護を考慮し，データの用途に合わせて個人情報を自動的に抽出するための手法について検討する．

以下では，まず第 2 節で，音声データの用途に応じて抽出すべき情報について述べる．また，個人情報を保護するために抽出すべき情報のうち，特に人名に焦点をあて，人名を抽出する手法について述べる．次に，我々の収集した音声データについて説明する．第 4.1 節ではいくつかの条件で行った書き起こしテキスト中の人名抽出実験と結果について説明し，第 5 節で実験結果に関して考察する．

2. 抽出すべき情報

我々は医療現場における事故の防止を目指し，看護師の行動や作業を理解するため，様々なウェアラブルセンサや環境センサを用いて医療現場にてデータ収集実験を行っている．本節では，看護師が装着したボイスレコーダによって収集した音声データについて述べる．

2.1 書き起こしテキストの用途

本研究の音声データの書き起こしテキストは以下のような用途に利用できると考えている．

- 看護師の行動分析，業務量調査

看護師の行動分析，業務量調査は，従来よりタイムスタディ調査として行われているもの [4] で，単位時間あたりの業務内容や業務転換数などを看護師の経験年数や対象患者の看護度に応じて分析し，看護師のスキル，業務量を評価するものである．その際に，患者や看護師の名前が具体的にデータ内にある必要はなく，看護師は経験年数などの看護経験によるラベリングがされていて，必要に応じて個人が特定できれば十分である．患者についても病症による看護度によりラベリングされているだけで十分な場合が多い．

- 対話研究

対話研究においては，話者の社会的な立場や話者間の関係が明確に把握でき，話者の置かれている状況が理解できれば，多くの場合，分析は可能である．また，我々のデータでは，医療事故に関係するデータ（インシデントレポート）も合わせて収集することを考えており，コミュニケーションエラーの原因究明も行えると考える．その分析のためには，時系列情報などの情報が必要である．

- 医療現場の用語調査

医療現場の用語調査では，カルテに書かれる病名や薬品名と実際に発話される病名や薬品名が違う場合が考えられる．例えば，患者に病名を伏せたい場合や薬品名が長いため略称が利用されているような場合である．この場合，個人名は特に必要なく，誰が誰に対して発話をしているときに，どのような用語を利用したかがわかれば良い．

以上 3 つのいずれの用途においても，患者や看護師の具体的な個人名が絶対に必要ということはない．従って，プライバシ保護の観点からも，少なくとも個人名は別の文字列に置き換える方が望ましい．この際に，抽出した個人名は何らかの方法でどういった役割の人物であるか区別できるラベルに置き換える必要がある．

プライバシ保護の研究と言えば，アクセスさせる人を限定する研究や改ざんを防ぐ研究がなされている [5]．我々の必要とするプライバシ保護は，データ全体の閲覧を禁止するのではなく部分的に情報を隠蔽削除し，個人識別が困難，あるいは，できないようにする形式のものである．

個人を特定できる情報としては，人名，年齢，住所，電話番号が一般的であるが，医療現場においては，住所，電話番号は書類に直接記入する場合が多く，対話中に詳細に現れることは少ない．また，年齢や電話番号は，部分的な表現や省略がされることが多く，単なる数字の列挙となることが多い．医療現場では薬品の量などを示す数字も多く発話されるため，その数字が何に関係するものか判断するのは難しい．

一方で，人名は，個人を特定するのに一番重要な情報であると共に，会話中では頻繁に出現する．また，話者の立場や話をしている状況によって，表現の仕方や出現パターンが多様であり，すべてを列挙することは困難である．しかしながら，医療現場における会話中の人名については，ある程度社会性のある場所で，話者同士の関係が見知らぬ人同士の会話であるため，「先生」「さん」や「さま」と言った接尾辞を伴って現れると考えられる．そのため，接尾辞を解析することで，人名をある程度判断することができると思われる．

その他，書き起こしテキストから，個人識別ができるのは，複数のタイプの情報が共起して出現する場合が考えられる．例えば，俳優の名前のように，映画タイトルと役名の情報が共起すると，誰であるか判断できるような場合や特殊な薬を投与されていると患者名が特定できてしまうような場合である．音声を収録した場所や状況にも依存すると考えられるが，少なくとも我々の対象とする医療現場での複数の情報，例えば薬品名と検査名，あるいは診療科などの組み合わせパターンとそれぞれの情報を抽出し必要に応じて削除する方法も検討する必要がある．

以上より，プライバシ保護の観点からは，最低でも人名を抽出し，何らかの処理を行う必要がある．また，看護師の行動分析のためには，個人識別に関係しそうな単語列，例えば薬品名，病名，診療科名，看護行為などを示す文字列も抽出する必要がある．さらに，事故防止に関係して，薬品の量を示す数字につ

いて抽出する必要がある．

なお，個人情報として話者の性別もあげられるが，会話中においては，音声を聞くだけで性別はほぼわかってしまう．また，話し方や声の質で個人を特定できることもある．本稿においては，書き起こしテキストに焦点をあて，上記のような音響的な特徴については取り扱わず，書き起こされたテキスト中の個人情報，特に人名について取り扱うこととする．

2.2 人名の抽出について

我々の収集した音声データには，ウェアラブル機器を持った看護師を中心とする会話が収集されている．看護師同士の会話，看護師と医師，患者，患者の家族と様々な対話が含まれている．また，実際の医療現場に於ける会話であるので，呼びかけをしたり，名前を名乗ることも多く，人名が頻繁に出現する．更に，収録現場は社会性のある場所なので，人名は「先生」「主任」や「さん」と言った接尾辞を伴って現れると考えられる．そのため，接尾辞を解析することで，人名とその人の社会的立場をある程度判断することができる．しかし，家族同士や仲間内での会話においては，通称，愛称，呼び捨てなどが存在する．また，人名に続く接尾辞についても「さーん」「ちゃーん」と言った長音を含む形のものなど，会話中に特徴的に現れる接尾辞が存在すると考えられる．

テキストからの人名の抽出技術として，固有表現抽出技術がある．MUC（Message Understanding Conference）において始まった固有表現抽出の研究[6]は，現在，様々な分野で利用され，拡張されてきている．特に，質問応答の技術開発のためには，質問の答えのタイプを認定し，そのタイプの情報を抽出する必要がある．応用範囲の拡大に従い，固有表現のタイプは拡張され，人名や組織名だけでなく，質問の分野における解答のタイプの数だけ固有表現のタイプを拡張され，整備されつつある[7]．また，固有表現抽出技術は新聞記事のような書き言葉を対象としていたものから，音声データ，話し言葉へも対象を広げ，研究されている[8]．しかしながら，既存の技術が話し言葉のテキスト，特に実世界の会話データに対しても精度良く利用可能なのか，詳細に分析されてはいない．更に我々の扱う領域に関しては，話し手の置かれている状況や話し手，聞き手の社会的関係により，人名の表現に多様性があると考えられるため，その多様な表現を持つ人名を高い再現率で抽出することが可能なのか調査し，確かめる必要がある．

3. 音声データ

本稿の実験で利用した音声データは，ある病院のある診療科で音声データ収集に特化したウェアラブル機（市販のICレコーダに一部改良を施した機材）で収録したものである．本収録においては，看護師の行動分析用と対話分析用の音声データ[9]を収録したが，本稿では誰の発話か比較的判断のしやすい行動分析用の音声データを利用した．

行動分析用の音声データは，看護師が何らかの業務やイベントを行う際に，ウェアラブル機付属のブザーを鳴らし，その業務やイベントを簡単な文章で発話する形で収録している．また，ブザーは10分間隔で自動的に鳴り，その時に看護師は何をやっていたか，簡単に作業行為について発話することになっているが，作業行為について発話できない場合は，その時の被験者の看護師と患者などの対話や環境音が自動的に録音される．そして，ブザー音を含む数秒間分の音声データを自動的に切り出し，切り出した音声データを各看護師毎に書き起こした．その書き起こしたデータが行動分析用の書き起こしテキストである．一例を表1にあげる．なお，書き起こし作業は，書き起こし作業経験者1名，薬学部出身者1名，元看護師（現場経験3年）1名，書き起こし作業初心者1名の4名を1グループとして行った．

表1 実験データ例

Table 1 Data Sample

Time	Utterance
11:28:11	オザクさんの点滴の準備を始めます．
11:32:01	オザクさんの点滴の準備終わりました．
11:32:04	サガラさんの血糖測ります．
11:33:48	サガラさん，血糖測り終わりました．
11:40:00	ナヤさーん，お薬置いておきますねー．

収録した音声データは，のべ39人，約500時間分のデータが得られており，そのうち，約50時間分の行動分析用音声データの書き起こし作業が終了している．本稿の実験では，この行動分析用テキストを利用した．このテキストの中には，個人名は2281個含まれていた．

本データの片方は，人名の漢字が不明確であることを想定し，個人名はカタカナ表記とした．一方，入院患者や病院関係者であれば，カルテや勤務管理表などから人名を前もって入手し辞書化することが可能であると考えられる．そこで，患者名，看護師名のリストを利用し，人手により正しい表記（漢字表記）への変更も行った．つまり，本稿の実験で利用する書き起こしテキストには，人名がカタカナ表記である場合，漢字表記の場合（人手変換）の2種類あり，それぞれを利用し，人名抽出実験を行う．なお，本稿中で例にあげる人名はプライバシを考慮し，実際の人名から別の人名に置き換えてある．

4. 人名抽出実験

固有表現抽出技術には，統計的手法とパターン照合に基づく手法とがある．統計的な手法では，珍しい名前や出現頻度の少ないものについては対応ができない．また，本実験データは接尾辞を調べることで人名をある程度把握できるだろうと予想されている．そこで，辞書やルールを用いたパターン照合に基づく固有表現抽出を行うこととした．固有表現抽出ツールとして，三重大学で開発されているツールNExTの0.82バージョンを利用した[10][11]．NExT利用時の形態素解析には茶筌2.3.3[12]を利用し，辞書はipadic2.7.0を利用した．

4.1 抽出実験1

まず手始めに，NExTや茶筌に変更を加えず，ネットから入手できる状態のままで抽出実験を行った．

書き起こしテキストは，看護師毎に個別のウェアラブル機を利用して収録し，それぞれ書き起こしている．そのため，抽出

結果は各看護師の書き起こしテキスト毎に算出した．表2は，個人名をカタカナで表記した場合，表3は入手した人名リストを参考に人手で漢字に変換した場合の人名抽出結果の一部である．全看護師のデータにおける再現率と精度の平均は，カタカナ表記の場合 **83.6%**，**86.5%**，漢字表記の場合 **89.2%**，**88.1%**であった．予想通り，漢字表記の方が良い結果となっている．

なお，再現率，精度については次の式を利用して算出した．

$$\text{再現率} = \frac{\text{システムが抽出した正解の数}}{\text{正解の数}} \times 100 \qquad (1)$$

$$\text{精度} = \frac{\text{システムが抽出した正解の数}}{\text{システムが人名として抽出した数}} \times 100 \qquad (2)$$

カタカナ表記の人名には，言いよどみ（コググレッさん）や長音（アーベーさーん）が含まれている．その表記を人手による漢字変換においては，正しい人名（小暮さん，阿部さーん）に変換されている．従って，これも結果に反映されている可能性がある．

言いよどみや長音の出現は看護師の話し方の特徴として偏りがあるか調べてみたところ，どの看護師にも，患者対応の仕事の場合には，2割程度出現率で存在し，看護師による出現の違いはみられなかった．

4.2 抽出実験 2

利用した抽出ツールは，新聞記事に対応したものであるため，書き言葉を対象としている．従って，実験1では，データ中の話し言葉の特徴を有する部分については抽出できなかった．実験2として，話し言葉でよく現れる「さーん」「ちゃーん」「くーん」「せんせー」の長音を含む表現も名詞の接尾辞として認識できるように抽出ツールの辞書を変更して実験を行った．利用したデータは，4.1節で利用したデータと同じものである．

実験結果の平均再現率と精度は，カタカナ表記の場合，**87.2%**，**87.0%**であり，オリジナル状態のツールの時と比較すると，平均再現率は4%ほど向上している．また，漢字表記の場合の平均再現率と精度は **92.5%**，**88.7%**であり，オリジナルの時と比較すると，平均再現率は3%ほど向上している．

4.3 抽出実験 3

実験2で，接尾辞についての話し言葉の特徴を有する部分の認識はできるようになったが，息継ぎなく発話されているような部分では句読点がないために，感動詞なども繋がっている単語列が人名と認識されている部分があった．例えば，「すいません桑原さん」や「あー納谷さん」といった表現である．NExTの人名に接続する単語に関するプログラムを修正し，感動詞は人名には含まれないように変更した．

平均再現率と精度は，カタカナ表記の場合，**90.4%**，**90.5%**であり，漢字表記の場合は **94.7%**，**91.2%**であった．オリジナルの時と比較すると平均再現率はカタカナ表記では7%，漢字表記では5%向上している．

5. 考　　察

5.1 抽出実験 1 について

抽出実験1の結果において，全く抽出できなかった人名は，

表4　抽出できなかった人名
Table 4　Non Extracted Results

	呼び捨て	未抽出	非パターン
カタカナ表記	105	4	76
漢字表記	59	11	74
	過長抽出	部分抽出	別分類
カタカナ表記	80	52	35
漢字表記	82	25	58

カタカナ表記の場合198個で，人名全体の約9%であった．漢字表記の場合，155個，約7%であった．以上のことより，NExTのツールを利用した場合，人名抽出の精度は人名表記が漢字でなくても9割以上であることがわかる．しかしながら，プライバシ保護を目的とする場合，抽出できていない残りの約1割分の人名を抽出する必要がある．また，データの利用目的に合わせ，人名を他のラベルに置き換える場合，同姓同名への対応も検討する必要がある．

まず，抽出実験1の結果において，抽出できなった人名，正しく抽出できなかった人名を分類すると，6つのグループに分けられた．全く抽出ができなかった3グループ（呼び捨て，未抽出，非パターン）と，人名を抽出できたが完全に正解と一致しなかった3グループ（過長抽出，部分抽出，別分類）に分けられる．その結果を表5(注1)に示す．

呼び捨て，未抽出，非パターンの3グループが，抽出ツールにより全く抽出されなかったグループである．呼び捨ては，文字通り，人名を呼び捨てている場合で，姓だけを連呼していたり，電話の対応において自分の名前を名乗ったときの表現である．非抽出は，人名と人名に繋がる接尾辞の形式にもかかわらず，抽出されなかった人名である．非パターンは，話し言葉に表れる接尾辞（「さーん」,「ちゃーん」など）を固有表現抽出用の接尾辞として登録していなかったために抽出されなかった人名である．

過長抽出，部分抽出，別分類は，固有表現として抽出はされたものの正解と完全に一致しなかったグループである．過長抽出は，人名の前の単語も含んで人名として抽出しているために正解と判定できなかった文字列である．例えば，「ちょっとクワハラさん」,「すみませんアベさん」といった文字列である．部分抽出は，人名の部分的なところのみ抽出している場合である．例えば，「オザクヒロミさん」のオザクだけを人名として抽出したり，「納谷さん」の谷だけを人名として抽出している場合である．また，別分類は，人名であるにも関わらず，地名や組織名として抽出された文字列である．

ツールにより抽出されなかった人名はカタカナから漢字に変換するだけで減少している．特に，呼び捨ての場合は，半減している．これは，ツールの辞書に人名として登録されている場合，その結果をうけて，接尾辞の形式に関わらず，抽出ツールが人名と判断しているためである．一方，未抽出の数は，漢字になると4個から11個と増加する．これは，ツールの辞書に

(注1)：この表において，抽出に失敗した人名の対応する特徴をすべてカウントしているため，抽出できなかった個数よりも多くなっている．

表 2　カタナカ表記データからの人名抽出結果の一部

Table 2　Personal Name Extraction Results from Katakana Form Data

看護師 ID	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	Ave.
個人名の数	13	16	25	26	30	34	43	46	61	67	108	115	132	145	245	58.5
抽出した数	17	12	27	25	31	37	42	39	61	60	105	113	139	147	248	56.5
正解した数	10	11	21	21	28	25	38	36	58	58	100	104	119	140	221	48.9
再現率 (%)	76.9	68.8	84.0	80.8	93.3	73.5	88.4	78.3	95.1	86.6	92.6	90.4	90.1	96.6	90.2	83.6
精度 (%)	58.8	91.7	77.8	84.0	90.3	67.6	90.5	92.3	95.1	96.7	95.2	92.0	85.6	95.2	89.1	86.5

表 3　漢字表記データからの人名抽出結果の一部

Table 3　Personal Name Extraction Results from Kanji Form Data

看護師 ID	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	Ave.
個人名の数	13	16	25	26	30	34	43	46	61	67	108	115	132	145	245	58.5
抽出した数	18	16	28	27	32	39	43	40	62	60	105	117	143	144	248	59.2
正解した数	11	15	22	23	29	28	38	37	60	58	102	107	125	137	224	52.2
再現率 (%)	84.6	93.8	88.0	88.5	96.7	82.4	88.4	80.4	98.4	86.6	94.4	93.0	94.7	94.5	91.4	89.2
精度 (%)	61.1	93.8	78.6	85.2	90.6	71.8	88.4	92.5	96.8	96.7	97.1	91.5	87.4	95.1	90.3	88.1

表 5　改良したツールで抽出できなかった人名

Table 5　Non Extracted Results

	呼び捨て	未抽出	非パターン
カタカナ表記	110	6	11
漢字表記	69	12	1
	過長抽出	部分抽出	別分類
カタカナ表記	12	125	56
漢字表記	13	30	44

は存在しない漢字の人名が，部分的に別の品詞に一致したり，未知語として分析されてしまうため，その結果，人名と判定できなかったものである．

非パターンに分類された人名の抽出については，話し言葉特有の固有表現キーを登録すれば対応できると考えられる．また，過長抽出は人名部分を確実に抽出しているため，プライバシ保護の観点からは特に問題がない．一方で，呼び捨て，未抽出，部分抽出，別分類については，何らかの対応をする必要がある．別分類として抽出される人名は辞書に地名や組織名として登録されており，その結果として，地名や組織名として抽出されている．そのため，辞書を変更する，あるいは人名に続く接尾辞を伴う名詞を強制的に人名と判断するなどの対応が必要と考えられる．さらに，呼び捨て，未抽出，部分抽出については，カルテ等から収集できる人名リストの利用を検討する必要がある．

5.2　抽出実験 2 と 3 についての考察

抽出実験その 2 とその 3 は，話し言葉特有の接尾辞や単語のつながりについて処理を可能とした場合の実験であった．

4.1 節の抽出実験の結果と比較すると再現率の平均はカタカナ表記の場合，83.6%から 90.4%へ，漢字表記の場合 89.2%から 94.7%へ上昇している．精度平均は，カタカナ表記の場合 86.6%から 90.5%へ，漢字表記の場合 88.1%から 91.2%へ上昇しているが，上昇幅が少ない．

現状で抽出できない人名は，人名として辞書に登録されていない漢字を含む人名で，なおかつ名詞以外の品詞に人名の一部の漢字が一致してしまうような場合である．例えば，「行○さん」，「渡○さん」，「○上さん」，「○下さん」といった人名である．特に「下さ」が動詞として認識されてしまうため，人名として抽出することができていない．また，看護師が患者の姓だけでなく名前も確認しているような時に，「小暮誰や」といった表現が現れる．この表現に対しては，人名の抽出ができない．この点については，話し言葉における人名の現れる場合の事例を集め，詳細にルール化するか統計処理をする必要がある．

プライバシ保護の観点においては，人名でないところが人名として抽出される場所があっても構わない．しかし，プライバシ情報を削除したデータを利用して他の分析を行う場合には，支障が出る可能性がある．そこで，人名でない単語であるが，人名として抽出された結果について考察する．

人名でないが，人名として抽出される単語列として，多い物は「患者さん」「ご本人さん」「おかゆさん」といった人名の接尾辞パターンを持っている単語列である．また，本データは話し言葉であるため，新聞記事とは違い，書き言葉での一般的な文章のような整った形になっていない文（単語列）が存在する．そのため，文末表現がない場合，部分的に辞書に一致する単語が人名として認識される．例えば，「ミドリン点眼（点眼の薬名）」のミドリを個人名としている場合や，「患者さん，のぶん，のー」というような言いよどみの部分では，のぶを個人名と抽出している．この結果については何らかのパターンがあると思われる．従って，より詳細な分析を行い，対処する必要がある．

抽出ツール内の辞書で，人名に関係する登録を削除し，接尾辞などの人名に関係するパターンをより詳細化して人名を抽出したところ，平均再現率がカタカナの場合 94.4%，漢字の場合 94.5%，平均精度はカタカナの場合 90.1%，漢字の場合 90.0%となった．本データには呼び捨ての部分が多く，人名を示すパターンを詳細化しても再現率はこれ以上あげることができないと考えられる．呼び捨ての出現パターンは，本実験データにおいては，電話での応対やカルテの確認をしている場合など，ある程度限定できる．音声以外のセンサデータを利用し状

況により，呼び捨てのパターンでも人名として抽出する方法を検討する必要がある．

人名である可能性の高い単語列，文字列を抽出すればシステムが抽出した人名候補、正解候補が増えるため，再現率は上昇すると考えられるが，本来，人名では無いものも抽出してしまうため，精度は下がる．地名と人名と組織名はそれぞれ同じ単語を使うことが多いことから人名が地名や組織名として抽出されるケースが多い．そのため，試しに，抽出ツールにおいて，地名，組織名として抽出された単語をすべて人名として精度を求めたところ，再現率は98.4%まであがったが，精度は20.4%にまで落ちてしまった．プライバシ保護の観点からは，より再現率を高くする必要があるが，再現率が高くなると，誤認識が増えるというトレードオフが生じる．特に，誤認識の部分が薬品名や検査名である場合，看護師の行動分析の観点からは支障を来す．さらに，対話分析においても，対話文中の多くが人名として削除されてしまっては，内容を把握することも難しく，支障を来すおそれがある．薬品名や検査名については，専門用語辞書を別途用意し，人名抽出とは別に取り扱う必要がある．

2.1節でも述べたが，我々の実験データは，実環境の対話データとして，対話構造の分析に利用できる．そのような分析の際には，患者や看護師，医師といった参加者の個人名は必要がなく，対話をしている話者の社会的関係，例えば，医師と患者，患者の家族と看護師といった枠組みが判別できればよい．また，発話者の年齢も大枠がわかれば良く体重や身長といった個人情報も不要である．その一方で，看護師の行動分析や看護記録作成支援においては，どのような病気の患者にどういった薬を投与したのか，個人名は不要であるが，年齢，体重，身長といった個人情報は，投与する薬の分量などに関係するため，必要となってくる．また，同姓の患者の扱いを間違えるなどのミスを防止するためには，同姓であったという情報は必要となる．実際，同姓で曖昧になる人名は，本実験を行った1週間の期間においてのべ162個，看護師，患者，主治医にも実在する同姓は1種類46個出現していた．このような曖昧性についてはカルテなどの医療現場で使われている情報を有効に利用することで解決できると考える．

現在では，手始めにプライバシ情報として，人名を抽出し削除，あるい別のラベルに変更することを検討しているが，データを分析するユーザの立場により抽出するべき情報は変化すると考えている．また，場合に応じて，特殊な病名や薬品名は個人を特定できるプライバシ情報となり得る可能性もある．プライバシ情報の選定と同時に，削除すべき情報のランク付け，評価方法なども検討する必要がある．

6. 終わりに

本稿では，我々のプロジェクトで収集している音声データから個人情報として人名を抽出する実験について報告した．今回，既存のツールを改良して実験を行った結果，94.7%まで，再現率を高くあげることができた．しかしながら，プライバシ保護の観点から考えると，100%の再現率が望まれる．今後のデータ分析利用のために，すべての人名を簡単に抽出できるよう，抽出支援ツールを開発する予定である．

一方で，人名抽出の再現率を高めれば，精度が下がり，本来人名ではない情報まで抽出してしまう．データの利用価値を下げずにプライバシ保護を確実に行う方法について，さらなる検討が必要である．また，データを利用する立場，目的に応じて，提示できる情報に変化を持たせる必要があり，プライバシ情報とデータ利用の目的と提示できるデータの関係を明確にする必要がある．そのため，我々の扱うドメインに合わせて個人名だけでなく，薬品名や病名なども抽出できるツールの開発が必要と考える．

行動分析用の音声データでは，口語的な特徴は長音，尻切れ文章の存在，呼び捨てであった．より自由な会話を対象にしている引き継ぎ時の音声データにおいても実験を行い，口語的な特徴を持つものが他にあるか調査する予定である．そして人名に関係するデータをさらに収集するとともに，対話コーパスでのプライバシ情報と分析情報の取り扱いについて検討し，看護師の行動分析，看護用語の収集，対話構造の解析も行う予定である．

謝辞

本実験に協力して頂いた医療機関の関係者の方々に感謝いたします．なお，本研究は通信情報研究機構 (NICT) の委託研究により実施したものである．

文　献

[1] L.T.Kohn, J.M.Corrlgan, M.S.Donaldson, “To Err Is Human:Building a Safer Health System”, The National Academies Press, Nov, 1999.

[2] 小暮潔, “E-ナイチンゲール・プロジェクト 日常行動・状況理解に基づく知識共有システムの構築に向けて”, 第7回知識科学シンポジウム,2005.

[3] NTCIR workshop, Cross-Language Question Answering Tasks, http://www.slt.atr.jp/CLQA/,

[4] 山本摂子，武田輝子，高橋孝子，天生目純子，長村聖子，須崎光子 “看護師の忙しさを構成しているもの—看護業務量調査の結果から—”, 第31回看護管理,pp.168-170, 2000.

[5] 本城信輔，洲崎誠一，齋藤司，三浦信治 “プライバシに配慮したWWWにおける個人属性認証・アクセス制御システム” 情報処理学会論文誌,Vol.43,No.8, 2002.

[6] R. Grishman, B.Sundheim, “Message Understanding Conference 6: A Brief Hsitory”, COLING, 1996.

[7] http://nlp.cs.nyu.edu/jneh/.

[8] F. Kubara, R. Schwarts, R. Stone, R Weischedel, “Named Entity Extraction from Speech”, in Proceedings of DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[9] 小作浩美, 相良かおる, 納谷太, 桑原教彰, 阿部明典, 小暮潔, “医療現場の音声収録とそのコーパス化”, 言語処理学会, 第11回年次大会,pp.958–961,2005.

[10] 桝井文人, 鈴木伸哉, 福本淳一, “テキスト処理のための固有表現抽出ツール NExT の開発”, 言語処理学会, 第8回年次大会,pp.176–179, 2002.

[11] 渡辺一郎, 桝井文人, 福本淳一, “固有表現抽出ツール NExT の精緻化とユーザビリティの向上”, 言語処理学会, 第10回年次大会,2004.

[12] http://chasen.naist.jp/hiki/ChaSen/