3Y-3

音声認識を用いた用例収集のためのプライバシフィルタリング手法の検討

東　拓央† 西村 竜一† 吉野　孝†
†和歌山大学システム工学部

1　はじめに

近年，訪日外国人者数の増加に伴い医療現場での多言語間コミュニケーションの機会が増えている．医療現場での多言語間コミュニケーションには，医療ミス等を未然に防ぐため正確性が求められる．様々な言語へと正確に翻訳された同じ意味の用例セットである用例対訳を用い，高精度な多言語間コミュニケーションを実現する研究が行われている [1] が，用例収集自体が容易ではないという問題がある．

そこでビデオチャットによる遠隔の通訳者を介した多言語間コミュニケーションの支援を行い，ビデオチャット使用時の会話を用例として収集する手法を提案する．本稿では収集された会話内容のプライバシ情報を保護するためのフィルタリング手法の検討を行う．

2　プライバシフィルタリングの必要性

図 1 に遠隔の通訳者を介した多言語間コミュニケーションにおける音声認識を用いた用例収集のイメージを示す．本研究では医療施設にいる診察や治療に訪れた外国人患者，医療従事者および外部医療通訳組織の医療通訳者の三者がビデオチャットを通じてコミュニケーションを行うことを想定している．このようなビデオチャットにおける三者間対話の音声を録音する．これにより医療時に必要となる用例や語句を自動的に取得することができる．取得した音声データは音声認識をかけてテキスト化し，用例収集サーバに蓄積する．蓄積された用例は他のシステムへの提供を考えているが，そこで問題になるのがプライバシ保護の問題である．本手法で集められた会話内容には個人のプライバシ情報が含まれると考えられる．他のシステムに提供可能な安全な用例とするためには，それらのプライバシ情報をフィルタリングする必要がある．

3　関連研究

本研究の関連研究として，土屋らの“プライバシ保護のための音声中の人名除去手法の検討”がある [2]．この研究では，ラベル無しコーパスと固有表現ラベル付きコーパスを併用することで，ラベル付きコーパスに頻出（存在）しない語を含む固有表現を頑健に抽出できる固有表現抽出法を利用し [3]，音声認識結果の文章に適用している．しかしこの手法では固有表現ラベル付きコーパスにとっての未知語には対応できるが，音声認識誤りによってうまくプライバシ保護することが難しいとされている．

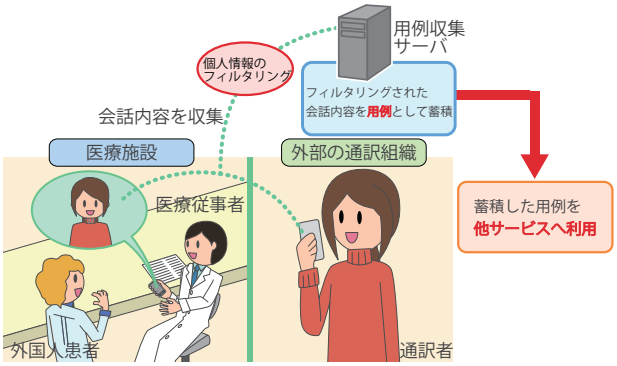

図 1: 遠隔の通訳者を介した多言語間コミュニケーションにおける音声認識を用いた用例収集のイメージ

4　音声認識済テキストのフィルタリング検討

4.1　音声認識済テキストの作成

新聞読み上げ音声コーパス (JNAS)[4] を用い，コーパス内の音声ファイルに音声認識をかけたテキストを作成し固有表現フィルタリングの検討を行った．音声認識には大語彙連続音声認識システム Julius[5] と Julius ディクテーションキット v.4.0 に付属している単語辞書 (約 6 万語の単語と音素列) を用いた．

4.2　フィルタリング手法の検討

4.2.1　CaboCha によるフィルタリング手法

CaboCha[6] は IREX[7] の定義による固有表現解析が可能な高性能な係り受け解析器である．今回は京都大学コーパス [8](Version 3.0，毎日新聞 95 年 1 月 1 日から 17 日までの約 4 万文) から学習済みの CaboCha を用い，抽出された固有表現に対しフィルタリングする手法を検討した．

CaboCha によるフィルタリングを行った際に特徴的であった例を表 1 に示す．この例では CaboCha が「金」という単語を「キム」と読んでしまったため，人物の名前として抽出してしまったと考えられる．この例のように漢字の読みによって単語の性質が変わってしまうものや，音声認識によってカタカナから漢字に変換されたものによって正しい単語の性質が取得できない現象が見られた．

4.2.2　音声認識単語辞書によるフィルタリング手法

大語彙連続音声認識システム Julius は認識対象とする単語と読み（音素列表記）を利用者自身で定義し，音

Study on Privacy Filtering Method for Collecting Parallel Texts using Voice Recognition
Takuhiro HIGASHI† Ryuichi NISIMURA† Takashi YOSHINO†
†Faculty of Systems Engineering, Wakayama University

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.

表 1: CaboCha によるフィルタリング例

オリジナル	カネができると次は名誉だがクリントン政権に加わったのはどうも本当にクリントンという指導者に米国の希望を見たためらしい
音声認識結果	金ができると、次は名誉だが、クリントン政権に加わったのは、どうも！本当にクリントンという指導者に、米国の希望を見たためらしい。
フィルタリング後	[人物] ができると、次は名誉だが、[人物] 政権に加わったのは、どうも！本当に [人物] という指導者に、[場所] の希望を見たためらしい。

表 2: 音声認識単語辞書によるフィルタリング例

オリジナル	カネができると次は名誉だがクリントン政権に加わったのはどうも本当にクリントンという指導者に米国の希望を見たためらしい
音声認識結果 (フィルタリング済)	金ができると、次は名誉だが、[人物] 政権に加わったのは、どうも！本当に [人物] という指導者に、[場所] の希望を見たためらしい。

声認識単語辞書を独自にカスタマイズすることが可能である．そこで，単語辞書内に登録されている固有表現の出力をあらかじめ固有表現とわからないように変更しておくことで，音声認識での固有表現認識時にプライバシ情報を隠蔽できるような手法を検討した．

音声認識単語辞書によるフィルタリング例を表 2 に示す．この例は CaboCha の手法で示した文と同じ文であるが,「金」という単語を正しく認識することができている．これは CaboCha が文脈のみで判断していたのに対し，Julius の単語辞書には読みの情報があり，音声が正しく入力されていれば読みの情報からも単語を判断することができるためであると考えられる．また，音声認識によって認識できる単語は単語辞書に存在するもののみであるため，認識できる固有表現をあらかじめ固有表現と認識できない形にしておくことで誤りなく固有表現がフィルタリングできる．そのため安全性は非常に高いといえる．

4.3 各手法のメリットとデメリット

今回検討した二つの手法についてメリットとデメリットを表 3 に示す．

日本語係り受け解析器 CaboCha によるフィルタリング手法では，音声認識エンジンの単語モデル自体を改変しなくてよいため汎用性があるが，音声認識後のテキストに適用するとフィルタリングに漏れが発生する可能性がある．また，音声認識単語辞書によるフィルタリング手法では，音声認識単語辞書内の固有表現の出力形を固有表現と認識できない形に編集しておくことによって安全なフィルタリングが可能であるが，音声認識エンジンの単語モデル自体をカスタマイズする必要があるため汎用的に扱うことができないという欠点がある．

表 3: 各手法のメリットとデメリット

	CaboCha によるフィルタリング手法	音声認識単語辞書によるフィルタリング手法
メリット	• 音声認識システム自体を改変する必要はないため汎用的に扱える	• 認識できる単語自体をコントロールできるため安全性は高い • 単語を判断する際読みの情報も利用できる
デメリット	• 音声認識結果に左右されやすい • 文脈のみでしか単語を判断できない	• 音声認識システム自体を改変するので汎用的には扱えない

5 まとめ

本稿では音声認識により出力されたテキストに適用するためのプライバシフィルタリング手法として，二つの手法の検討を行った．

CaboCha によるフィルタリング手法では汎用性が高い反面，フィルタリング漏れが発生する場合がある．また，音声認識単語辞書によるフィルタリング手法は安全性が高い反面，汎用性に乏しいという欠点があることがわかった．

本研究では医療分野という個人情報の安全性を重視しなければならない限られた分野を対象とするため，音声認識単語辞書によるフィルタリング手法でフィルタリングを行っていく予定である．今後はビデオチャットのシステムにフィルタリング手法を組み込み，効果を検証する．

参考文献

[1] 宮部真衣ほか：外国人患者のための用例対訳を用いた多言語医療受付支援システムの構築，電子情報通信学会論文誌. D，情報・システム J92-D(6)，pp.708-718(2009).
[2] 土屋雅稔ほか：プライバシ保護のための音声中の人名除去手法の検討，言語処理学会，第 16 回年次大会，PA2-31(2010).
[3] Tsuchiya,M.，Hida,S，and Nakagawa,S.：Robust extraction of named entity including unfamiliar word，Proceedings of ACL-08:HLT，Short Papers(Companion Volume)，pp.125-128(2008).
[4] 伊藤克亘ほか：大語彙連続音声認識研究用日本語コーパス：JNAS，Journal of the Acoustical Society of Japan (E) 20(3)，pp.199-206(1999).
[5] 李晃伸：大語彙連続音声認識エンジン Julius ver.4，電子情報通信学会技術研究報告．SP，音声 107(406)，pp.307-312(2007).
[6] 工藤拓，松本裕治：チャンキングの段階適用による日本語係り受け解析，情報処理学会論文誌 43(6)，pp.1834-1842(2002).
[7] 関根聡，伊佐原均：IREX：情報検索、情報抽出コンテスト，情報処理学会研究報告．自然言語処理研究会報告 98(82)，pp.109-116(1998).
[8] 黒橋禎夫，長尾眞：京都大学テキストコーパス・プロジェクト，言語処理学会，第 3 回年次大会発表論文集，pp.115-118(1997).

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.