1K3-OS-17a-4

軽度認知症高齢者を対象とした情報支援システムにおける言語知識を用いない自由発話の意図推定

Sub-lexical Dialogue Act Classification of Spontaneous Speech in an Information Support System for the Elderly with Mild Dementia

佐土原 健[*1] Ken Sadohara　児島 宏明[*1] Hiroaki Kojima　成田 拓也[*2] Takuya Narita　二瓶 美里[*2] Misato Nihei　鎌田 実[*2] Minoru Kamata　大中 慎一[*3] Shinichi Onaka
藤田 善弘[*3] Yoshihiro Fujita　井上 剛伸[*4] Takenobu Inoue

[*1]産業技術総合研究所 National Institute of Advanced Industrial Science and Technology (AIST)
[*2]東京大学 The University of Tokyo
[*3]日本電気株式会社 NEC Corporation
[*4]国立障害者リハビリテーションセンター研究所 Research Institute of National Rehabilitation Center for Persons with Disabilities

This paper concerns the dialogue act classification in a spoken dialogue system that delivers necessary information to the elderly with mild dementia. Although lexical features have been shown to be effective for the classification, automatic transcription of the spontaneous speech requires elaborate costly language modeling. This paper therefore considers a classifier not requiring language modeling that uses sub-lexical features instead of the lexical features. It operates on sequences of phonemes obtained by a phoneme recognizer and exhaustively analyzes the saliency of all possible sub-sequences by using a support vector machine with a string kernel. An empirical study on a dialogue corpus of elderly speech shows that the sub-lexical classifier performs better than a lexical classifier using hidden Markov models of words, and is is robust against the case where the mismatch of the language model is large.

1. 序論

高齢化社会の到来に伴い認知症者の数は増加しており，特に，症状が比較的軽い段階で維持される軽度認知症者が増加すると言われている．軽度認知症者が日常生活を送る上で最も大きな問題となるのは，記憶障害と注意障害であり，日付や時間，その日の予定などの情報の欠落が，自立した行動を妨げ，円滑な日常活動を困難にする．そこで，こうした困難を緩和するために，記憶や見当識を補助する情報支援機器が有効と考えられており，デイプランナー（スケジュールや時間把握），アラーム付き薬入れ（服薬時間の支援）などが開発されている．本研究では，支援する情報の種類毎に異なるデバイスを用いる必要のない汎用的な情報支援システムの開発を目標としている．ただし，システムの使用方法の習得という負担を利用者にかけないよう，家族や介助者が日常的に行っている方法と同様に，対話を通して必要な情報を支援できるシステムの開発を目指している [Inoue 12].

このような音声対話を可能にするためには，対話システムが利用者の発話意図を推定し適切な応答を行う必要がある．対話における発話の意味は談話行為と呼ばれ，発話に対して，あらかじめ定められた談話行為タグを付与する談話行為識別は対話処理における最初の重要なステップと考えられており，これまでに多くの研究がなされてきた．これらの研究により，談話行為識別において最も効果的な素性は，単語やフレーズといった語彙的な素性であることが明らかになっている．音声対話において，このような語彙素性を利用するためには，音声をテキスト化する必要があるが，日常会話のようなくだけた自由発話の音声認識には音響的，言語的なモデリングに多くの困難が伴う [Furui 05]. 本稿では，特に自由発話における言語モデリングの困難さに注目する．

自由発話においては，フィラー，言い淀みや言い直し等が頻出するが，こうした表現の出現パターンは話者により大きく異なることが知られている．このような意味的に冗長な表現は，談話行為識別のような言語処理に悪影響を及ぼすことから，文献 [Honal 05] では，冗長表現除去のために話者毎に異なるモデルを用いている．また，冗長な表現に限らず，語彙の使用に制約の少ない自由発話においては，話者毎に言語モデルを適応することが有効である [Nanjo 04]．しかし，そのような話者適応を行うためには，話者毎に多くの言語資源を必要とするので，我々の対話システムに適用するのは困難である．

さらに，自由発話においては，発音の変動が大きく，辞書の適応も必要になる場合がある．文献 [Akita 10] では，一つの単語に対して，実際の発音に対応する複数のエントリを導入することで音声認識精度が向上することを示した．この結果が示すように，正確なテキスト化には，実際の発音に合わせた辞書の適応が必要になるが，一つの形態素に対して，複数の表現形が存在することは，談話行為識別のような言語処理に悪影響を及ぼす．従って，認識されたテキストの正規化が必要になるが，日本語の単語は，実際の発音に合わせて複数の表現形を持ちうるので，英語におけるステミングのような単純な正規化を行うことは困難である．

以上，述べたように，語彙素性の利用には正確なテキスト化が必要で，そのためには，コストの大きい高度な言語モデリングが必要になる一方で，談話行為識別という目的に限ってみると，正規化を行わない正確なテキスト化は必ずしも有用ではないことが分かる．そこで本研究では，語彙素性のよりコストの低い利用法として，音声を音素列として認識した上で，単語の代わりに音素の部分列を素性とする談話行為識別手法について考察する．この手法では，もしある単語が談話行為識別に有効であれば，その単語の断片，つまり，その単語を構成する部分音素列の一部は，同様に識別に有効であることを仮定している．もし，その仮定が正しければ，効果的な言語モデリングが困難で，識別に重要な単語が誤認識されるような状況であって

連絡先: 佐土原 健, ken.sadohara [at] aist.go.jp

表 1: 談話行為タグとその出現割合．二人のラベラーの一致率は 81.9%, $\kappa = 0.782$.

タグ	例	%
情報要求	晩ご飯は何を食べましたか？	0.2
確認要求	わかりましたか？	8.2
行為要求	トイレに行ってみたらどうかな？	4.3
注目要求	○○さん。ちょっといいですか？	15.1
言い直し要求	えっ？なんだって？	2.1
肯定的返事	はい。わかりましたよ。	26.9
否定的返事	いいえ。	0.2
情報提供	魚を食べたよ。	60.0
あいさつ	ありがとうね。	15.1
注目表示 (承認)	そうね。	19.9
注目表示 (不承認)	ほんとう？	0.2
その他	笑い声等	5.4

も，その単語の断片 (部分音素列) が保存され，談話行為識別に寄与することが期待できる．しかも，音素列を利用することで，自由発話に顕著な発音の変動を織り込み，さらには誤認識の傾向さえも反映した識別が可能になるかもしれない．

本稿は，以下のように構成される．まず，次節で，我々の情報支援システムのための談話行為タグ集合を導入した後，3. 節で，従来からある単語ベースの談話行為識別器と本研究が提案する音素ベースの談話行為識別器を示す．次に，4. 節において，軽度認知症高齢者の協力を得て作成した対話コーパスを用いて，二つの識別器の性能を比較検討する．

2. 情報支援システムのための談話行為タグ

本研究では，家族や介助者が日常的に行っている方法と同様に，対話を通して，自立した生活行動に必要な情報を適切なタイミングで支援することが可能な汎用情報支援システムの開発を目指している．このシステムは，軽度の記憶障害と注意障害がある場合でも情報の伝達を可能にするために，以下のような手順で対話を行う: (1) 注意喚起，(2) 先行連鎖 (情報が伝達される旨を伝える)，(3) 情報伝達, (4) 対話終了．各対話状態では，利用者が対話に追従できているかどうかを必要に応じて確認し，必要があれば同じ対話状態を繰り返すこともある．

このような対話状態の遷移を計算機上で実行する目的で，利用者の発話意図を区別する 12 個の談話行為タグ (表 1) を用意した．談話行為タグは，談話における発話の機能を表したタグであり，対話管理，要約，曖昧性解消や音声認識といった多様な音声言語情報処理において有用であると考えられている．領域非依存なタグセットに関する研究 [荒木 99] がある一方で，領域やタスクに合わせて独自のタグセットを用いることも多い．本研究でも，開発する情報支援システムに合わせた上記タグセットを設計した．

本研究では，NEC 社製の PaPeRo© を用いて，プロトタイプシステムを構築し，有料老人ホームにて独居生活をしている 20 名の被験者の協力を得て，居室かそれに近い環境の一室において，実験者が補助しつつロボットと被験者一人づつとの音声対話の収録を行った．被験者の内，3 名は男性で残りの 17 名は女性であり，平均年齢は 82.9 ± 7.2 (67 歳から 97 歳)，MMSE スコアは 21.4 ± 5.8 (9 点から 30 点) であった．得られた 7,123 発話の書き起こしに加えて，談話行為タグの付与を 2 人のラベラーで独立に行ったところ，タグの一致率は 81.9% で，コーエンの κ は 0.782 であった．

3. 談話行為識別

談話行為の自動識別については，これまでに多くの研究がなされているが，素性と識別モデルの 2 つが識別器の設計における重要な論点となる．素性については，用いられている語彙，統語情報，韻律情報，談話構造など様々な情報源に基づいた素性が提案されている．本研究では，語彙 (lexical) 素性の代わりに，音素列のような sub-lexical な素性を用いる談話行為識別について考察する．その際，システムの発話の談話行為を文脈情報として同時に利用する．

識別モデルについては，隠れマルコフモデル (HMM), 最大エントロピーモデル，条件付き確率場，サポートベクトルマシン (SVM) など様々なモデル化手法が提案されている．本研究では，任意の部分音素列の識別への寄与を網羅的に分析する目的で，音素列上で動作する文字列カーネルを用いた SVM で識別器をモデル化する．このような識別器について記述する前に，文献 [Stolcke 00] で用いられている，HMM を用いた単語ベースの識別器を, 我々の問題に特化した形で以下に示す.

3.1 HMM を用いた単語ベース識別器

文献 [Stolcke 00] では，発話 E_i が, 観測できない談話行為 U_i に依存して生起するとし，U_i にマルコフ性を仮定することで，最適な談話行為系列 U^* を以下のように定式化した．

$$U^* = \underset{U}{\operatorname{argmax}} \prod_{i=1}^{n} P(U_i|U_{i-1})P(E_i|U_i).$$

本研究では，直前の発話はシステム側の発話であり，その談話行為 U_R は観測可能である仮定し，利用者の発話 E の最適な談話行為 U^* を以下のように求める．

$$U^* = \underset{U}{\operatorname{argmax}}\, P(U|U_R)P(E|U).$$

発話が単語系列 $W_1, \ldots, W_n$ として観測されるときには，独立同分布を仮定する．

$$U^* = \underset{U}{\operatorname{argmax}}\, P(U|U_R) \prod_{j=1}^{n} P(W_j|U)$$

発話が音響特徴量 A として観測されるときには，音声認識システムが出力する N-best 解 $W^{(n)}$ $(1 \le n \le N)$ を用いて，以下のように，U^* を求める．

$$\begin{aligned} U^* &= \underset{U}{\operatorname{argmax}}\, P(U|U_R)P(A|U) \\ &= \underset{U}{\operatorname{argmax}}\, P(U|U_R) \sum_{n}^{N} P(A|U, W^{(n)})P(W^{(n)}|U) \\ &= \underset{U}{\operatorname{argmax}}\, P(U|U_R) \sum_{n}^{N} P(A|W^{(n)})P(W^{(n)}|U), \end{aligned}$$

ここで，最後の等式は，$P(A)$ が $W^{(n)}$ だけに依存していることを仮定しているが，もちろんこれは一般には成立せず，U は，$W^{(n)}$ の発音様式に影響を与えることに注意されたい．また，$P(A|W^{(n)})$ には，音声認識システムが計算する音響尤度を用いることができるが，大語彙の認識システムではこの値が非常に小さいので，アンダーフローを避けるために，N-best 解の最大尤度 $M = \max_n P(A|W^{(n)})$ を用いて以下のように

計算する．

$$U^* = \operatorname*{argmax}_U \frac{P(U|U_R)}{M} \sum_n^N P(A|W^{(n)})P(W^{(n)}|U)$$
$$= \operatorname*{argmax}_U P(U|U_R) \sum_n^N \exp(L(n)) \quad (1)$$
$$L(n) = \ln(P(A|W^{(n)})) - \ln(M) + \ln(P(W^{(n)}|U))$$

3.2 SVM を用いた音素ベース識別器

本研究では，対話音声を音素認識した上で，音素列間の類似性に基づいて談話行為を識別する手法の可能性を検討する．音素列間の類似性として，文字列カーネル [H.Lodhi 02] を用い, 識別学習アルゴリズムとしては SVM を用いる．

文字列カーネルは，ある文字列に含まれる任意の部分文字列の出現頻度を成分とする特徴ベクトルの内積として定義される．その際，音素の挿入誤りや脱落誤りに対処するために，部分文字列のギャップの個数 g に応じた減衰因子 λ^g $(0 < \lambda \leq 1)$ を課して頻度を数える．また，音素の置換誤りに対処するために，異なる音素であっても音素間の類似性行列に基づいて応分の出現頻度を与えることが出来る．上記のようなギャップを含んだ部分文字列の出現頻度に基づく特徴ベクトルの内積は，動的計画法を用いることで，任意の文字列 s, t と部分文字列の長さ p に対して，$O(p|s||t|)$ の計算量で計算することが可能である．このような音素列に特化した文字列カーネルについて，詳細は，文献 [Sadohara 10] を参照されたい．ただし，本研究では，システム側の発話の談話行為タグを文脈とする識別を行うため，以下のような拡張を行う．今，任意の音素列を s, t とし，各々の音素列の文脈を $c(s), c(t)$ とするとき，s と t の類似性 K を以下のように定義する

$$K_\ell(s,t) \stackrel{\text{def}}{=} \delta_{c(s),c(t)}\kappa_\ell(s,t),$$

ここで，$\delta_{C(s),C(t)}$ は，クロネッカーのデルタであり，κ_ℓ は文献 [Sadohara 10] で定義された, 長さ ℓ の部分文字列に対する文字列カーネルである．

さらに，文字列カーネルは，部分文字列の長さ ℓ 毎に計算されるカーネル関数の重み付け和に拡張される．

$$K^{\leq p}(s,t) \stackrel{\text{def}}{=} \sum_{\ell=1}^{p} \gamma_\ell K_\ell(s,t).$$

このようなカーネル関数は，実際に, ある特徴空間における内積であることが分かるので Mercer 条件を満たしている．

以上，文字列カーネルと SVM を用いて，音素列がある談話行為であるか否かを分類する方法について述べたが，複数の談話行為タグを識別するには多クラス分類を行う必要がある．本質的に 2 クラス分類器である SVM を多クラス分類に拡張するさまざまな試みが行われているが，本稿では，単純に，1 対他方式を採用する．すなわち，談話行為タグ U 毎に，U であるか否かを分類する 2 クラス分類器 f_U を学習し，任意の音素列 s に対して，$U^* = \operatorname{argmax}_U f_U(s)$. さらに，式 (1) と同様に，音素認識エンジンの N-best 解を用いる場合，

$$U^* = \operatorname*{argmax}_U \sum_n^N P(A|s_n) f_U(s_n)$$
$$= \operatorname*{argmax}_U \sum_n^N \exp(\ln(P(A|s_n)) - \ln(M)) f_U(s_n)$$

のように，最適な談話行為を計算する．

表 2: 談話行為識別性能.

	word-HMM		phone-SVM	
	Accuracy	F1	Accuracy	F1
Transcript	0.800	0.521	0.817	0.624
ASR	0.758	0.521	0.789	0.563

4. 実験

音素ベースの談話行為識別器の有効性を評価するために，上記音声対話コーパスを用いて，被験者の 4,080 発話の談話行為を識別する実験を行った．この実験では，被験者毎の複数の実験日の内，後半の実験から得られた 2,160 発話を評価データとし，前半の 1,920 発話を学習用データとした．その際，被験者の発話にほとんど現れなかった行為要求, 注目要求, 否定的返事，注目表示 (不承認) を除いた 8 種類の識別を行った．

識別実験には，人手で書き起こしたテキストと音声認識システムの出力を用いた．音声認識エンジンには，Julius [Lee 01] を用い，辞書と単語トライグラムは，上記学習データから作成した．辞書のエントリー数は 1,008 で評価データに対する未知語率は 6.37%, パープレキシティは 17.97 であった．音響モデルは，連続音声認識コンソーシアムが配布している高齢者用の性別非依存モデル (3000 状態 64 混合 PTM triphone) [Baba 01] をベースとして, 学習用データを使って MLLR により話者適応を行ったものを使用した．このような音声認識システムを用いた単語正解精度は 42% であった．同様な音声認識システムを用いて連続音素認識を行い音素書き起こしも作成した．ただし，このシステムでは，単語トライグラムの代わりに音素トライグラムを学習用データから作成し使用している．音素の正解精度は，54% であった．

音声認識結果から談話行為識別器を学習する際は，認識エンジンの 5-best 解全てを学習データとして用いた．一方，評価時には，前述したように，10-best 解を音響尤度で重み付けして談話行為タグを予測した．SVM に用いたパラメータは，$\lambda = 0.7$, $\gamma_\ell = 1.0$ を用いた．部分文字列の最大長 p については，辞書中の単語の平均音素長が 4.8 音素であったことから $p = 4$ とした．さらに，ソフトマージンパラメータ C は $C = 10.0$ としている．

表 2 は実験結果をまとめた表である．ここで，'accuracy' は，談話行為タグ予測の正解率を表し，'F1' は，F-measure, すなわち，再現率と適合率の調和平均を各タグ毎に計算した平均を表している．この表からわかるように，音素ベースの識別器 (phone-SVM) は単語ベースの識別器 (word-HMM) よりも識別性能が高く，書き起こし (Transcript) 対しては，有意水準 5% (McNemar 検定による) で，音声認識結果 (ASR) に対しては，有意水準 1% で有意に高性能である．以下，さらに詳細にこの結果を分析する．

4.1 音素ベース識別器の頑健性

図 1 は，二つの識別器の正解率を示しているが，書き起こしと音声認識結果に対する正解率に加えて，評価データも使って作成した辞書と言語モデルを用いた認識結果 ASR(CHEAT) に対する正解率が示されている．この図からわかるように，単語ベース識別器は，ASR と ASR(CHEAT) との間に有意な性能差がない．これは，学習データの量を増やし言語モデルの適合性を向上させたとしても，正解率は向上しないことを示唆している．一方で，言語モデルのミスマッチが大きい新たな利用者が現れた場合，識別性能が劣化することは十分に起こりうる．

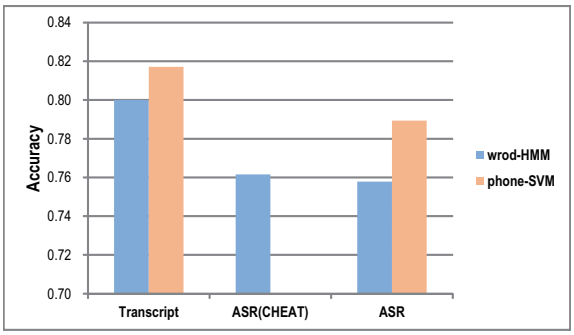

図 1: 単語ベース識別器と音素ベース識別器の比較

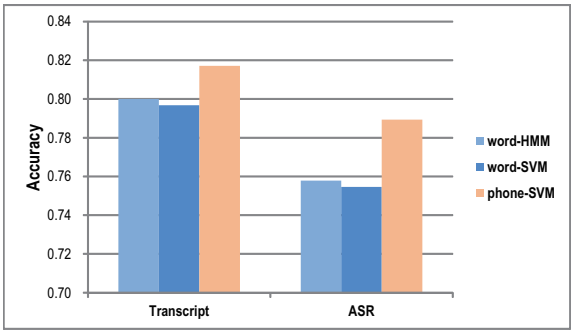

図 2: 単語素性と音素素性の比較.

その場合でも，言語知識に依存しない音素ベースの識別器では性能劣化は生じ得ないので，この意味で音素ベース識別器は言語モデルの不適合に対して頑健であると言える．

4.2 音素素性の有効性

図 2 は二つの識別器の正解率に加えて，単語素性 (bag-of-words) を用いた SVM (word-SVM) の正解率が示されている．この図からわかるように，word-SVM は word-HMM と同程度の識別性能しかない．従って，phone-SVM と word-HMM の性能差は，SVM を用いたモデル化手法の差に起因するのではなく，単語素性と音素素性との差に起因することが分かる．

特に，書き起こしに対する phone-SVM と word-SVM の性能差は，音声認識とは無関係に純粋に単語素性と音素素性の違いによる．これは，一つの形態素に対して，発音に忠実に書き起こすことで生じた複数の表記により word-SVM の性能が劣化しているためではないかと考えられる．さらに，この性能差は，音声認識おいて，さらに拡大している．双方とも同一の音声認識エンジンと音響モデルを用いた認識結果であることを考えると，その理由は，word-SVM において，言語モデルの不適合により識別に有用な単語素性が失われる場合であっても，phone-SVM においては，それら単語の断片（音素部分列）が残っており，それを手がかりに正しい識別が行われているためではないかと考えられる．

5. 結論と今後の課題

本研究は，軽度認知症高齢者を対象とした音声対話システムのサブシステムとして，言語知識を用いない音素ベースの談話行為識別器を提案した．この識別器は，自由発話を正確にテキスト化するために必要な高コストの言語モデル作成を行うことなく，連続音素認識システムが出力する音素列から談話行為を識別することができる．それを可能にするために，通常の単語ベースの識別器が行っている，識別に有用な単語の分析の代わりに，識別に有用な単語の断片（部分音素列）の分析を行う．そのような分析を網羅的かつ効率良く行うために，音素列上で動作する文字列カーネルを用いた SVM を用いており，音素の挿入，脱落，置換誤りが生じる場合であっても，識別に有用な部分音素列に基づいて談話行為タグを予測することができる．このような識別器の有効性を，軽度認知症高齢者の音声対話コーパスを用いて評価し，単語ベースの識別器が言語モデルの不適合により十分な識別性能を達成できないときでも，音素ベースの識別器はより高い識別性能を達成可能であることを確認した．今後は，本手法を他の多くの研究で用いられているコーパスに適用し性能を比較することの他に，韻律素性を取り込んだ識別器について検討を進めていきたい．

謝辞

本研究にご協力頂いた (株) 生活科学運営と各施設の入居者の皆様に謝意を表する．本研究の一部は，JST の戦略的イノベーション創出推進プログラムの支援を受けて実施した．

参考文献

[Akita 10] Akita, Y. et al.: Statistical transformation of language and pronunciation models for spontaneous speech recognition, *IEEE TASLP*, Vol. 18, No. 6, pp.1539–1549 (2010)

[Lee 01] Lee, A. et al.: Julius — an open source real-time large vocabulary recognition engine, in *Proc. of Eurospeech*, pp. 1691–1694 (2001)

[Baba 01] Baba, A. et al.: Elderly acoustic model for large vocabulary continuous speech recognition, in *Proc. of Eurospeech*, pp. 1657–1660 (2001)

[Furui 05] Furui, S.: Spontaneous speech recognition and summarization, in *Proc. of HLT*, pp. 39–50 (2005)

[H.Lodhi 02] Lodhi, H. et al.: Text classification using string kernels, *JMLR*, Vol. 2, pp. 419–444 (2002)

[Honal 05] Honal, M. et al.: Automatic disfluency removal on recognized spontaneous speech - rapid adaptation to speaker-dependent disfluencies, in *ICASSP* (2005)

[Inoue 12] Inoue, T. et al.: Field-based development of an information support robot for persons with dementia, *Technology and Disability*, Vol. 24, No. 4 (2012)

[Nanjo 04] Nanjo, H. et al.: Language model and speaking rate adaptation for spontaneous presentation speech recognition, *IEEE TSAP*, Vol. 12, No. 4 (2004)

[Sadohara 10] Sadohara, K.: Kernel topic segmentation for informal multi-party meetings and performance degradation caused by insufficient lexicon, in *Proc. of SLT*, pp. 430–435 (2010)

[Stolcke 00] Stolcke, A. et al.: Dialogue act modeling for automatic tagging and recognition of conversational speech, *Computational Linguistics*, Vol. 26, No. 3 (2000)

[荒木 99] 荒木 雅弘 他 : 発話単位タグ標準化案の作成, 人工知能学会学会誌, Vol. 15, No. 2, pp. 251–260 (1999)