ワードグラフを考慮したシステム要求検出*

◎佐古淳，山形知行，滝口哲也，有木康雄 (神戸大)

1 はじめに

近年，音声による機器操作インターフェイスが実用化されつつある．特に，ロボットとのコミュニケーションや，カーナビの操作など，手を使うことが困難な機器の操作への適用がなされている．音声インターフェイスを用いる際，システムへの要求発話とそれ以外の発話を区別する必要がある．このため，物理的なスイッチを用いることで区別を行ったり，ネットワーク文法を用いて自動的に区別を行う手法 [1] が提案されている．また，我々も音声認識結果に対してブースティングを用いることで，システム要求発話の柔軟性・多様性を保持したまま自動的に区別を行う手法の提案を行ってきた [2]．ただし，音声認識結果を用いることから，認識誤りがシステム要求の検出に悪影響を与えるという問題があった．

本研究では，音声認識とシステム要求検出をワードグラフ上で統合的に行うことで，システム要求検出性能の向上を試みる．このとき，以下のような二通りの定式化が可能である．ひとつは，発話の目的（システム要求かそれ以外か）に応じた言語モデルを用いる手法．もうひとつは，認識の仮説パスに対して識別的に発話の目的を推定する手法である．まず，通常の音声認識器を用いてワードグラフを構築し，そのワードグラフ上で上記の２つの手法を実装し，実験を行った．また，実験は，従来の認識結果に対しブースティングを行う手法，システム要求発話のみで trigram を構築し，認識結果の信頼度を用いる手法についても行い，比較した．以下，次章において手法の詳細について述べる．

2 提案手法

観測信号系列を $\mathbf{O} = (o_1, \cdots, o_t)$，単語列を $\mathbf{W} = (w_1, \cdots, w_n)$，システム要求か否かを $s \in (0, 1)$ とすると，システム要求検出を同時に行う音声認識は以下のように定式化できる．

$$\begin{aligned}(\hat{s}, \hat{\mathbf{W}}) &= \underset{(s,\mathbf{W})}{\operatorname{argmax}} P(s, \mathbf{W}|\mathbf{O}) \\ &= \underset{(s,\mathbf{W})}{\operatorname{argmax}} P(\mathbf{O})^{-1} P(s, \mathbf{W}, \mathbf{O})\end{aligned}$$

ここで，$P(s, \mathbf{W}, \mathbf{O})$ をベイズの定理により，以下の二通りに展開できる．

$$P(s, \mathbf{W}, \mathbf{O}) = P(s) \cdot P(\mathbf{W}|s) \cdot P(\mathbf{O}|\mathbf{W}, s) \quad (1)$$

$$P(s, \mathbf{W}, \mathbf{O}) = P(\mathbf{W}) \cdot P(\mathbf{O}|\mathbf{W}) \cdot P(s|\mathbf{W}, \mathbf{O}) \quad (2)$$

式 1 の定式化は，言語モデル・音響モデルが s に依存するようなモデルを用いる手法となる．ただし，本研究では，音響モデルの s への依存は無視し，s に依存する言語モデルのみを考慮した．すなわち，

$$P(\mathbf{W}|s) = \prod_i P(w_i|w_{i-1}, \cdots, w_{i-N+1}, s) \quad (3)$$

という s に依存した N-gram を言語モデルとして用いた．

式 2 の定式化は，言語モデル・音響モデルは通常のものを用いる．加えて，認識仮説 $\mathbf{W}$ や観測音声 $\mathbf{O}$ から直接 s を推定するような確率モデルが存在する．このモデルについて，本研究では，ブースティングを用いた従来手法を sigmoid 関数を用いて擬似確率化して用いた．すなわち，弱識別器の数を T 個，弱識別器を $f_t(\mathbf{X})$，弱識別器の重みを α_t とし，

$$P(s|\mathbf{W}, \mathbf{O}) = \frac{1}{1 + \exp(-w_1 \sum_t \alpha_t f_t(\mathbf{W}, \mathbf{O}) - w_0)}$$

を用いた．ここで，w_1 及び w_0 は重み係数であり学習により推定する．これにより，音声認識とブースティングによる識別を統合的に行った．次節で，ここで用いたブースティングアルゴリズムについて述べる．

2.1 ブースティングによるシステム要求検出

本節では，ブースティングを用いて識別的にシステム要求か否かの判別を行う手法について述べる．本研究では，ブースティング法として Schapire ら [3] の提案している AdaBoost を用いた．

ブースティングは，複数の識別器による重み付き投票のための学習アルゴリズムである．学習によって，投票を行う識別器とその重みを決定する．投票による識別では，各識別器が相補的な関係にある方が性能が向上するとされる．ブースティングは，相補的な識別器を効率的に構成するためのアルゴリズムである．

以下に，AdaBoost 法の学習アルゴリズムを簡潔に述べる．まず，学習サンプルに対して均等な重みを与える．次に，最初の弱識別器 $f_t(t = 1)$ を選択し，弱識別器の重みを得る．このとき，識別を誤る学習サンプルの重みの合計が最も小さくなるような弱識別器を選択する．最後に，正しく識別された学習サンプルの重みを下げ，識別を誤った学習サンプルの重みを上げる．これにより，前の識別器とは傾向の異なる識別器が新たに選択され，相補的な識別器を構成できる．これらを T 回繰り返すことで，投票に用いる弱識別器 f_t と投票重み α_t を得る．

弱識別器には，Decision Stumps を用いた．これは，テキスト中の “素性” の有無によって識別を行う単純な手法である．例えば,「ください」があればシステム要求である，と識別する．このような単純な識別器をいくつも組み合わせることで高精度な識別器が構

*System Request Detection Based on Word-Graph, by Atsushi SAKO, Tomoyuki YAMAGATA, Tetsuya TAKIGUCHI and Yasuo ARIKI (Kobe University)

成できる．本研究では，unigram 及び bigram を素性として用いた．音響による素性は用いなかった．

3　システム要求判別コーパス

本研究で用いたシステム要求検出タスクについて述べる．本タスクでは，まず，二人以上の人間とシステムが同時に存在することを想定する．これは，ロボットを操作する際に周囲に人がいる場合や，カーナビを操作する際に助手席に同乗者がいる場合のように，自然な状況であると考えられる．二人以上の人間が互いに会話を行いながら，任意にシステムへの要求発話を行う．本研究では，“システム” として，音声により移動，写真撮影などの動作を行うロボットを用いた．典型的な利用方法としては，少し離れた場所から「こっちに来て」とロボットを呼ぶ,「写真を撮って」と写真を撮ってもらう，などがある．収録は，二人の発話者それぞれの胸元に取り付けたマイクで行った．発話数は 330，内 52 発話がシステム要求発話であった．

4　実験

本章では，提案手法を用いたシステム要求検出実験について述べる．まず，通常の音声認識器によりワードグラフを構築し，その上で 2 種類の提案手法を用いた認識，及びシステム要求か否かの識別を行った．評価は，システム要求検出の再現率・適合率によって行った．また，比較手法として，認識結果に対してブースティングを行う手法，システム要求発話のみで trigram を構築し，認識を行った際の信頼度によってシステム要求か否かを識別する手法についても実験を行った．

4.1　音声認識条件

音響モデルは，まず，CSJ モニター版のうち男性話者 200 名の講演音声を用いて作成し，これにテストセット・クローズドな話者適応を行ったものを用いた．適応データの分量は，約 10 分であった．言語モデルは，実験で用いた発話を書き起こしたテキストから作成した．ただし，テストセットに対してオープンとなるように，話者 B の発話のみを用いて話者 A の認識用言語モデルを作成した．通常の音声認識実験の結果，単語正解精度は 42.1%であった．このような音声認識条件の下でワードグラフを構築し，実験に用いた．

4.2　システム要求検出結果

前節の条件で得られたワードグラフに対し，提案手法によるシステム要求検出を行った．学習とテストは 10 folds のクロスバリデーション法により行った．素性は unigram + bi-gram を用いた．ブースティングの学習は，誤りを含む認識結果を用いて学習を行った．システム要求か否かに依存した言語モデルを用いた手法を “WG N-gram”，ブースティングによる識別結果を sigmoid により疑似確率化した手法を “WG Boosting”，従来の認識結果を用いたブースティングによる手法を “Boosting”，trigram の信頼度を用いる手法を “Confidence” とする．実験結果を図 1 に示す．

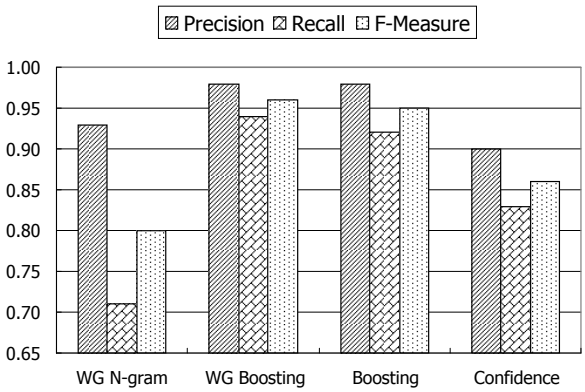

Fig. 1　Results of system request detection.

実験の結果，提案手法のうち，ブースティングの結果を sigmoid により疑似確率化して用いる手法が最も良い性能を示した．次いで，認識結果を用いたブースティングが良い性能を示した．元々ブースティングによる識別が高性能な上に，提案手法では，音声認識誤りを原因とする識別誤りが，ワードグラフ中から正しい単語を拾ってくることで解決し，性能が向上したものと考えられる．一方，システム要求か否かに依存する言語モデルを用いる手法や信頼度を用いる手法では，性能が低下した．特に,「こっちに来て，とか」のように「とか」という 1 単語の有無によって結果が左右されるような発話の識別を多く誤った．

5　おわりに

本稿では，システム要求検出を音声認識をワードグラフ上で同時に行う手法について述べた．二通りの定式化を行い，それぞれの手法及び，比較手法について実験を行った．実験の結果，提案手法のうち，ブースティングの結果を sigmoid により疑似確率化して用いる手法が最も良い性能を示した．認識結果が誤っている場合に，ワードグラフ中から正しい単語を拾うことで識別性能が向上したものと考えられる．一方，システム要求か否かに依存した言語モデルを用いる手法では，高い性能を示すことはできなかった．本タスクでは，ひとつの単語の有無により結果が変わってしまうことがあるため，N-gram はそのような識別には不向きであると考えられる．

今後の課題として，大規模なコーパスを構築し実験を行うこと，多様な表現によりシステム要求が可能なタスクにおいて実験を行うことがあげられる．

参考文献

[1] 石塚, SP98-5, Apr. 1998.
[2] 佐古, 音響論（春）, pp. 22-23, 2007.
[3] R.Schapire, Annals of Statistics, vol.26, no.5, pp.1651–1686, Oct. 1998.