

話題連想辞書の適用による携帯電話音声認識率の向上*

津金 ミキ, 唐澤 博†

山梨大学 工学部‡

E-mail: {tsugane, karasawa}@jewel.yamanashi.ac.jp

1 はじめに

音声認識システムにおいて、誤った単語の挿入や正しい単語の脱落などの誤認識が起こることは避けられない。本研究は誤認識に対して誤り訂正は行わず、音声認識ソフトによって与えられるスコアと話題連想辞書を用いて確からしい認識部分を抽出し、必須格については推論をし解析を行うことで認識率の向上を図る。

2 システムの概要

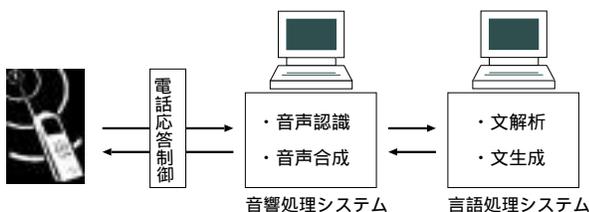


図1. 対話システム

図1のシステムは筆者等が取り組んでいる人工知能システムGAIAプロジェクトの“プロダクト”携帯電話を移動端末とした音声対話人工知能システムTel-GAIA”の開発を目的としている。

確らしい単語の抽出や必須格の推定は、言語処理部の解析段階で行う。

2.1 携帯電話音声の認識状況

携帯電話音声の平均認識率は、約90%（電波受信状態が最良の場合）である。

表1. 使用機器

PC	IBM NetVista
電話応答ボード	システムソフィア VCS3200
音声認識ソフト	IBM ViaVoice98
携帯電話	NTT DoCoMo N502it

*Improvement in the rate of cellular-phone speech recognition by application of a topic associative dictionary

†Miki Tsugane, Hiroshi Karasawa

‡Yamanashi University, 4-3-11 Takeda, Kofu, Yamanashi 400-8511, Japan

3 音声認識手法

3.1 確からしい単語の抽出

音声認識ソフトによって認識単語のスコアが得られる。そのスコアと話題連想辞書を用いて確らしい認識部分を抽出単語と決定し、それ以外の単語は切り捨てる。

不明確な部分は次の発声で復元できるチャンスがあるので、確からしい部分から文解析していき、対話を進めていく方法が有効だと考える。

3.1.1 スコアの特徴

スコアは認識された単語の確率の相対的な尺度で、-100 ~ 100の範囲で表される。よって、スコアは確からしい単語と不明確な単語の判断基準となるといえるので、スコアを本手法に取り入れた。

図2、表2に以下の文章のスコア付けした認識例を記した。

磐田が残り1分の劇的VゴールでJリーグ史上初の両ステージ制覇を達成した。第2ステージ優勝に王手をかけた磐田は東京V戦で延長後半14分、MF福西崇史(26)がVゴールを決め、1-0で勝ち、第1ステージに続く優勝を決めた。

磐田(-10)が(10)残り(-4)一分(12)の(12)劇的(-5)V(-15)ゴール(3)で(6)Jリーグ(-2)非常(-20)初(-16)の(8)上(-6)ステージ(-13)制覇(3)を(-3)達成(8)した(14)第二(-3)ステージ(1)優勝(14)に(13)王手(-1)を(2)かけた(10)磐田(-4)は(6)東京(0)ヴェルディ(1)戦(1)で(1)延長(4)後半(29)十(-18)四分(20)ミッドフィルダー(1)福西(-11)尚(-7)が(1)V(-11)ゴール(-1)を(-3)決め(7)いきたい(-13)中(-18)で(0)勝ち(-8)第(-2)一(-2)ステージ(1)に(-1)続く(6)優勝(2)を(-2)決めた(7)

図2. 認識結果 ()内: 前単語のスコア

表2をみると、スコアの値がマイナスでも正しく認識しているものが数多くあり、低いスコアを持つ単語を誤認識単語と決めてしまうことはできない。

3.1.2 話題連想辞書の適用

話題連想辞書とは、筆者等の研究室で開発したもので、話題ごとに共起している単語群が集められた辞書である。これによって、複数の単語から話題を決定した

表2. 認識単語とそのスコア(昇順) *: 誤認識単語

*非常	-20	第二	-3	優勝	2
十	-18	を	-3	ゴール	3
*中	-18	第	-2	延長	4
初	-16	一	-2	で	6
V	-15	を	-2	は	6
ステージ	-13	Jリーグ	-2	続く	6
*いきたい	-13	王手	-1	決め	7
一分	-12	ゴール	-1	決めた	7
福西	-11	に	-1	の	8
V	-11	東京	0	達成	8
磐田	-10	で	0	が	10
勝ち	-8	ステージ	1	かけた	10
*尚	-7	ヴェルディ	1	の	12
*上	-6	戦	1	に	13
劇的	-5	で	1	した	14
残り	-4	ミッドフィルダー	1	優勝	14
磐田	-4	が	1	四分	20
制覇	-3	ステージ	1	後半	29
を	-3	を	2		

り、話題から連想される単語群を引き出すことができる。

よって、スコアが低い単語でも文脈的に正しい単語(話題に関連ある単語)を残す方法として話題連想辞書を適用することを考えた。

3.1.3 スコアと話題連想辞書を用いた単語抽出法

1. スコアによる抽出単語の決定
スコアが0未満の単語を切り捨てる候補とし、0以上のものは抽出単語とする。
2. 話題連想辞書による抽出単語の決定
切り捨て候補単語が抽出単語の属する話題の要素であるとき、それを抽出単語とする。
3. 抽出単語以外の単語の切り捨て
上記の抽出単語以外を切り捨て単語と決定し切り捨てる。

3.2 話題連想辞書の概要

3.2.1 扱う話題

扱う話題は「サッカー」「野球」「ラグビー」「バレーボール」「柔道」「相撲」の6話題である。

3.2.2 話題の共起語

話題に関連深い単語(共起語)群を選出する。その選出方法は、各話題について書かれたデータを話題ごとに収集し、そのデータに出現する単語を抽出し、それら

の単語に重み付けをし、重みの高い単語を共起語とする。

1. 収集データ
ウェブ報知 [1] から上記6話題について書かれた記事を利用した。
2. 名詞抽出
データからもっとも話題を特徴づけると考えられる名詞を抽出対象とした。しかし、名詞の中でも話題を特徴づけるとは考えにくいものが複数存在するので、名詞の中でも一般名詞、固有名詞、サ変名詞の3つを抽出する。また、未知語のほとんどが固有名詞なので、未知語についても抽出対象とする。
3. 単語の重み付け
単語の重み付けにはTF/IDF法を用いた。

$$w_t^d = tf(t, d) * idf(t)$$

$tf(t, d)$: 話題 d 中に出現する単語 t の頻度

$idf(t)$: 全話題中に出現する単語 t の話題出現頻度

4. 共起語の決定
単語の重み付け結果をもとに閾値を設定し、閾値以上に位置する単語を話題の共起語とする。

表3. 各話題の共起語数

サッカー	143	バレーボール	91
野球	112	柔道	102
ラグビー	89	相撲	110

3.2.3 連想システム WAVE の適用

決定した共起語を連想システム WAVE[2]を用いて相互結合させ、話題から連想される単語群を引き出せるようにする。

4 おわりに

本研究では扱う話題数が少ないため、今後、広い話題に対して対話が行えるように話題数を増やしていく。

参考文献

- [1] 小川, 中村, 遠藤 ほか: ウェブ報知(オンライン), <http://www.yomiuri.co.jp/hochi>, (参照 2002-11-25)
- [2] 角田, 田中: PDAI&CD に基づく意味の学習および文脈依存の多義性解消, 電子情報通信学会技術研究報告, Vol.DE93-1, pp.1-8, 1993