E-014

誤認識を含む音声認識テキストからの連想検索によるキーワード抽出の検討
Keyword Mining from Speech Recognition Text

前岡 淳† 木村 淳一†
Jun Maeoka Junichi Kimura

1. はじめに

音声認識技術は音声コマンドや議事録作成，会議要約作成といった様々な用途での実用化が見込まれている．音声認識技術の適用先として，発話内容を音声認識し，話題の内容を反映したキーワードを抽出することで，Web 検索のキーワード入力を支援する機能や，発話内容に対するタグ付けを自動化する機能といった用途が考えられる．発話内容からのキーワード抽出において課題となるのは，音声認識の認識精度である．近年，音声認識技術の精度は飛躍的に向上したものの，100%の認識は不可能であり，誤認識が発生することを前提として利用する必要がある[1].

そこで，本研究では，誤りの含まれるテキスト文の中から内容を表すキーワードを抽出する手法として，概念検索や類似文書検索に利用される連想検索技術[2][3][4]を適用する．連想検索技術とは，類似の文書を高速に検索する技術であり，キーとなる入力文書に含まれる単語頻度に基づいて，類似する文書をスコア付けして検索することが可能である．誤りを含む音声認識テキストに対して，連想検索によって音声認識テキストに類似の文書群を検索し，それらの類似文書の中からキーワードを抽出することで，元の音声認識テキストに含まれる誤りの影響を低くすることが期待できる．

このような背景から，連想検索技術に基づくキーワード抽出手法について，各種サンプル音声入力の音声認識結果テキストを用いたキーワード抽出実験を行い，抽出精度評価を行った．

2. 提案手法

2.1 連想検索

連想検索技術は，多数の文書データベースから，ある入力文書に類似した文書を高速に検索する用途に用いられる．連想検索技術では，入力する文書(類似した文書を探したい文書)と，検索対象となる文書データベースのそれぞれの文書を，文書に含まれる各単語の頻度ベクトル同士を比較し，類似スコアを算出することで，文書間の類似性を判断する．連想検索を用いることで，検索したい文書テキスト全文をそのまま検索エンジンに投入すること可能となり，入力テキスト文書の内容から，検索のための単語選択を人手で行う必要がなくなる．

連想検索技術は，以下の二つの検索アルゴリズムを組み合わせることで類似文書の検索を実施する．

(1) 文書から単語の連想

指定した入力文書(群)から，それらに関連度の高い単語群をスコア付けし，検索するアルゴリズム

(2) 単語から文書の連想

指定した入力単語(群)から，それらに関連度の高い文書群をスコア付けし，検索するアルゴリズム

本研究では，音声認識結果テキストに類似した文書の抽出に単語→文書連想，抽出した類似文書群からの特徴単語抽出に文書→単語連想を用いる．

† 日立製作所 中央研究所 Central Research Laboratory, Hitachi, Ltd.

2.2 キーワード抽出アルゴリズム

図 1 に音声からキーワードを抽出する提案手法の流れを示す．

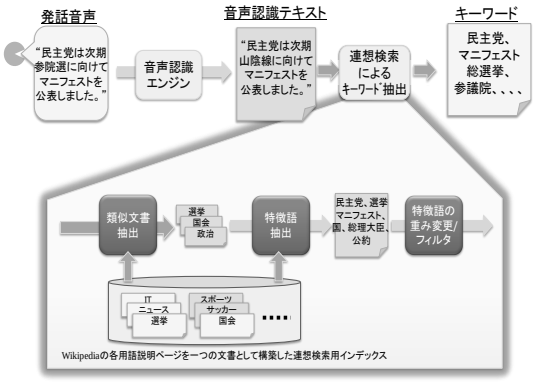

図 1 音声からのキーワード抽出

連想検索を用いたキーワード抽出アルゴリズムについて順を追って説明する．

(1) Wikipedia 全文からの連想検索用インデックスの構築

今回の実験では，発話内容に類似した文書を検索するための検索対象の文書群として，以下の観点から， フリー百科事典 Wikipedia の全文データ[5]を用いた．

- 様々な分野の語句を幅広く，平均的な分量でカバーしている．
- 最新の用語を含み，頻繁にアップデートがなされている．
- GNU Free Documentation License (GFDL)に基づき，全文データがダウンロード可能である[6].

Wikipedia の各語句に関する説明ページのテキスト内容を，連想検索における“一文書”ととらえ，各文書の平文テキストから，forward index および inverted index からなる連想検索用インデックス(以下，単にインデックスと呼ぶ)[3]を構築する．以下，インデックスの構築手順を示す．

a) テキスト化

配布されている Wikipedia 全文データは，XML によって構造化されているため，XML タグを排除したテキスト文に変換する．

b) 表現の正規化

Wikipedia 全文テキストに対して，表現を正規化する．半角カタカナ→全角カタカナ，全角数字→半角数字，全角アルファベット→半角アルファベットにそれぞれ統一する．

c) 頻度ファイルの作成

Wikipedia テキスト文のそれぞれの語句説明ページを，形態素解析プログラム茶筌[7]を用いて形態素に分割し，出現するすべての名詞の出現頻度を集計した頻度ファイルを作成する．

d) インデックスの構築

全ての語句説明ページに関する頻度ファイルから，連想検索用インデックスのフォーマットに変換する．

以上のインデックスの構築を，(2)以降のキーワード抽出処理に先だって実施しておく．なお，構築したインデックスは，文書数 702,384，異なり単語数 802,512 である．

(2) 音声認識

音声認識エンジンによって発話音声ファイルから認識文(テキスト)に変換する．音声認識エンジンには，大語彙連続音声認識エンジン Julius[8]を用いた．

(3) 認識文に対する頻度ファイル作成

認識文を茶筌により形態素に分割し，認識文中の単語出現回数を集計した頻度ファイルを作成する．このとき，連想検索用インデックスに存在しない単語は，類似文書抽出の算出には用いられない．

(4) 類似文書の検索

認識文の頻度ファイルを入力として，単語→文書の連想検索を実施し，類似文書群を抽出する．ここで抽出された文書は，認識文に類似した文書，すなわち同様の話題の文書であると仮定できる．実験では，抽出文書数を類似スコアの上位 100 文書とした．文書数を多くするほど，類似度の低い文章が含まれるようになり，後段の特徴語抽出処理において連想度の低い単語が抽出される率が高まる．

(5) 類似文書中の特徴語の抽出

次に，抽出した文書群を入力として，文書→単語の連想検索を施し，特徴語とスコア(以下，連想検索スコアと呼ぶ)を抽出する[4]．抽出するキーワードの数は特徴度の高いスコアから順に，(6)の選定を行ったうえで決定する．

(6) キーワードの選定

(5)で抽出した特徴語から，以下の重み付け，フィルタを施したものを最終的なキーワードとして選定する．

- 一文字のワードの排除
- 認識文中の出現頻度に応じた重み
- 固有名詞の重み

キーワード選定の際のスコアの定義を以下に示す．

キーワード選定スコア
＝ 連想検索スコア×認識文出現重み×固有名詞重み
連想検索スコア： 連想検索により抽出した特徴語のスコア
認識文出現重み： 1.0＋認識文の出現回数×0.1
固有名詞重み： 固有名詞ならば 2，それ以外ならば 1

実験では，キーワードスコアが 2.0 以上の特徴語を最終的なキーワードとして抽出とした．したがって，入力音声に対する抽出キーワード数は可変である．

3. 実験

ニュース文，ブログ文，日常会話を著者が読み上げた 30 のサンプル音声に対して，提案手法によるキーワード抽出を行った．サンプル音声から抽出したキーワードに対して，①発話内容を表す/要約するものであるか，②発話内容から連想し，気付きにつながるものであるか，の判断基準から，適合するか否かを主観評価し，抽出キーワードの適合率として算出した．また，音声認識の認識率 100%を想定するために，サンプル音声の正解文テキストについても，同様の手法でキーワード抽出を行った．

図 2 に音声認識率とキーワードの適合率の集計結果を示す．

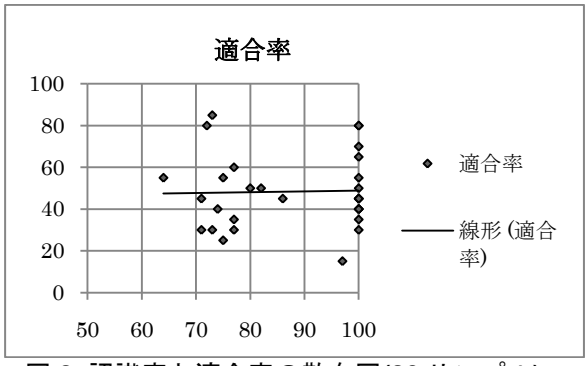

図 2 認識率と適合率の散布図(30 サンプル)

実験結果から，認識率 100%のサンプルの適合率の平均は約 51%，それ以外のサンプルの適合率の平均は約 46%であった．提案手法では，約 65%～の認識率の音声認識文に対して，認識率の低下に伴う適合率の著しい低下を伴うことなく，約 45%以上の適合率が得られた．今後はより多くのサンプル文に対する評価を継続し，評価精度を高めていく予定である．

4. おわりに

本研究では，誤認識が含まれる音声認識結果テキストからキーワード抽出する方法として，誤り単語によるキーワード抽出への影響を少なくすることを目的として，連想検索技術を適用した手法を提案し，実験を行った．その結果，提案手法により，音声認識の認識率が約 65～100%の音声認識文について，約 45%以上の適合率でキーワードを抽出できる見込みを得た．

今後は，実験サンプル数増やすとともに，話題に応じた適合率の変化について評価することで，定量的評価精度を向上させた上で，提案手法によるキーワード抽出の適合率向上を図る．

参考文献

[1] 古井貞熙，“人と対話するコンピュータを創っています 音声認識の最前線”，角川学芸出版，2009.
[2] 高野明彦他，“汎用連想計算エンジンの開発と大規模文書分析への応用”，第 19 回 IPA 技術発表会，2000.
[3] 安田知弘他，“連想検索エンジンのスケーラビリティおよび障害耐性の向上”，情報処理学会第 69 回全国大会，2007.
[4] 大規模連想検索エンジン MANTA，日立 http://chishiki.t.u-tokyo.ac.jp/event/20060706/niwa.pdf
[5] Wikipedia:データベースダウンロード http://ja.wikipedia.org/wiki/Wikipedia:データベースダウンロード
[6] Wikipedia:著作権 http://ja.wikipedia.org/wiki/Wikipedia:著作権
[7] 形態素解析器茶筌 http://chasen-legacy.sourceforge.jp/
[8] Linux 版 Julius ディクテーション実行キット v4.0(Julius-4.1.3) http://julius.sourceforge.jp/