情報処理学会
グループウェアとネットワークサービス研究会
2005年11月10日, 11日

Referred Conference Paper

モバイル環境においてニュース記事を音声検索するための認識語彙抽出手法の検討

荒金 陽助† 塩野入 理† 金井 敦†

† 日本電信電話株式会社，NTT 情報流通プラットフォーム研究所

あらまし 携帯電話をはじめとした様々な機器の発達により，モバイル環境におけるネットワークサービスが広く使われるようになってきた．このようなネットワークサービスの一つに，ニュース情報の提供がある．本稿では，このようなサービスにおけるニュース検索を音声認識インタフェースで行うための，検索キーワード用音声認識語彙抽出手法について検討を行った．高雑音環境下での音声認識では，音声認識率に認識語彙数の与える影響が大きいと言われている．そこで，利用者が発声すると考える語句の抽出と共に，単語の出現頻度などを用いた語彙数削減手法について提案した．従来本分野で用いられている手法に TF-IDF 法があるが，被験者実験による評価を行うことで提案手法と TF-IDF 法との比較を行った..

キーワード 時事キーワード抽出, 複合名詞, モバイル環境, 語彙数削減, 出現頻度

A Study for News Keywords Extraction Method for Mobile Conditions

Yosuke Aragane† Osamu Shionoiri† Atsushi Kanai†

† NTT Information Sharing Platform Laboratories, NTT Corporation

Abstract According to development of mobile technology, we are able to use several network services in the movile environments. In those services, there is a news information providing service. In this paper, we propose a keywords extraction method for news articles retrieval by voice recognition system. In high noisy environments such as mobile environment, the ratio of voice recognition succeedings are greatly influenced by a number of vocabulary for voice recognition. Therefore, the proposal method not only extract available keywords, but also cut the number of vocablary by using the appearance frequency of each keyword. In this field, TF-IDF method is one of a well-known technique. We have experimental evaluation for availability our method in comparison with TF-IDF technique.

Keywords Keywords extraction, Compound nouns, Mobile environments, Decrease of recognition vocabulary, Appearance frequency

1 はじめに

モバイル機器の発達・普及につれてモバイル環境向けの様々なネットワークサービスが提供されるようになってきた．特に携帯電話を対象として様々なネットワークサービスが提供されている．携帯電話向けの Web メニューでは，携帯電話の特徴である着信メロディや待受画面を始めとして，乗換案内やショッピング情報など PC 向けのポータルサイトが提供している情報種別の多くが利用可能となっている．本稿ではこれらのサービスの中でも，日々情報が更新され定期的な利用が想定されているニュースを対象として，モバイル環境で本サービスを利用するためのインタフェースについて議論する．

ニュースサービスの利用形態としては，(1) サイトを表示，(2) 見出し文一覧を表示，(3) 読みたい見出しを選択して本文を閲覧，というプロセスが一般的であると思われる．しかしながら，(2) の見出し

文を表示する際に画面サイズの制約から同時に表示可能な見出し文が限定されたり，自分の閲覧したいニュースを検索することが困難であったりといったインタフェースの課題があると考えられる．一般的に，オフィスやホームと比較して過酷な入出力環境となるモバイル環境でのインタフェースに関しては，様々な検討が行われている [1, 2]．そのようなインタフェースの中で，音声認識インタフェースは手操作を要求しないという特徴を持つことから，カーナビゲーション装置やボイスポータルシステムなどにおいて多く利用されている [3, 4]．そこで本稿では，音声認識インタフェースを用いてニュースを検索する利用形態について検討することとする．

一般のテキストベースのインタネット検索エンジンなどでは膨大な個数の単語を抽出し，それを対象として部分一致などの文字検索を行うことで検索結果を該当させ，検索率の向上を図っている．しかし，音声認識では，音声認識率と認識対象語彙数がトレードオフの関係にあり，実用的なサービスを考えるならば，認識語彙はあまり増やせない．さらに，認識対象語彙において部分一致よりも完全一致で認識を行う方が処理が容易なため，認識対象語彙を正確に設定する必要がある．

語彙数削減という課題に対して，従来よく用いられている手法に TF-IDF 法がある [5]．TF-IDF 法は，あるキーワードの文書内出現回数である TF (Term Frequency) と，全文書数におけるあるキーワードが出現する文書数の割合の逆数である IDF (Inverse Document Frequency) の積である．しかし，TF-IDF 法は構成要素の特性を反映していないなどの不十分な点が指摘されている [6]．そこで本稿では，音声認識インタフェースを用いてニュース記事を検索するための，ニュース記事の特徴であるニュースカテゴリを利用した認識語彙抽出法について提案する．そして，被験者による評価を通して提案手法の有効性を検証する．

2 提案手法

本稿で検討するインタフェースでは，ユーザの発声したキーワードが音声認識語彙として登録されている必要があり，ユーザが発声すると考えられる多くのキーワードをニュース記事から抽出することが重要となる．一方，音声認識語彙数と音声認識率とはトレードオフの関係にある．「過去一定期間のニュース記事を検索する」といったネットワークサービスでは，100 以上の記事を対象に検索を行うこととなる．従って，重複のために線形でないにしろ認識語彙が増加することは避けられない．音声認識率はサービス性に直結するため，トレードオフ関係にある音声認識語彙数の削減は非常に重要となる．従って，ユーザが発声する可能性の低いキーワードについては可能な限り認識語彙から排除することが必要となる．本章では，提案する “キーワード抽出手法” および “キーワード削減手法” についてそれぞれ説明する．

2.1 キーワード抽出手法

一般的にキーワードとしては名詞が用いられる．そこで提案手法としては，ニュース記事の見出し文および記事本文を形態素解析にかけ，抽出した名詞を元にキーワードを構成する手法を採用した．なお，本稿では形態素解析機として茶筌 [7] を利用した．

しかしながら，形態素解析は文章を最小の形態素に分割することが目的であるため，ユーザの発声するキーワードとは乖離してしまうことが多い．音声認識では処理を容易にするために，認識語彙と発声されたキーワードの完全一致によって認識処理を行うことが多いため，部分一致では合致したと見なされないこととなる．従って，ユーザの発声したキーワードの中に形態素解析によって抽出された短い名詞が含まれていたとしても，音声認識処理において “合致” とは見なされないことが多い．**表 1** にユーザの発声に対して，認識語彙が細分化されてしまった例を示す．

表 1 に示す形態素解析機によって抽出された名詞は断片的であり，音声認識によるニュース記事検索サービスにおいて，ユーザが発声するとは考えづらい．そこで本稿では，「ユーザは，複数の名詞からなるある意味を持ったキーワードを発声することが多いだろう」という仮定を置き，名詞が連続したものを “複合名詞” として認識語彙に採用することを考える．表 1 の例では，“複合名詞化” によって「部分開業」「九州新幹線」「丹波町」「鳥インフルエンザ」「イラク基本法」「署名拒否」「イラク復興支援会議」「長嶋監督」という認識語彙が抽出されることとなる．

また，認識語彙抽出処理の対象として記事見出し文を採用することが多い．しかし，記事見出し文で

表 1: 形態素解析結果とユーザ発声の例

文章	抽出された名詞	ユーザ発声例
部分開業する九州新幹線	部分, 開業, 九州, 新幹線	九州新幹線
丹波町の鳥インフルエンザ	丹波, 町, 鳥, インフルエンザ	鳥インフルエンザ
イラク基本法への署名拒否	イラク, 基本, 法, 署名, 拒否	イラク基本法
マドリードでのイラク復興支援会議	マドリード, イラク, 復興, 支援, 会議	イラク復興支援会議
長嶋監督の入院	長嶋, 監督, 入院	長嶋監督

はニュース特有の短縮された言い回しや，記事内容を抽象化した言葉での表現が多く，ユーザが発声するキーワードとは必ずしも合致しないことが考えられる．そこで本稿では，認識語彙抽出処理の対象を記事本文まで広げ，その複合名詞を認識語彙として採用する手法を提案する．

2.2 キーワード削減手法

本節では，音声認識率とトレードオフの関係にある認識語彙数の削減のための手法について説明する．

2.2.1 カテゴリ毎の出現頻度による削減

検索とは多くの情報から知りたい情報を絞り込む作業である．従って，ユーザは知りたい情報を特徴づけるキーワードを発声することが多いと考えられる．様々な記事に出現するキーワードによって検索を行った場合，検索結果として合致するニュース記事数が多くなり，検索を行う意味が無くなってしまう．そこで,「様々な記事に出現するキーワードは発声されることが少ない」という仮定を考え，認識語彙から削除する手法を検討する．

ここで,「様々な記事に出現する」ことを定義する必要がある．ニュース記事は社会や経済，国際，スポーツなどのカテゴリに分類されており，それぞれのカテゴリ内では特徴的なキーワードが多用されることがあっても，カテゴリを超えて特徴的なキーワードが多用されることは少ないと考えた．そこで，あるキーワードの各カテゴリにおける出現回数の標準偏差を利用した出現頻度閾値法を提案する．あるキーワード k のカテゴリ i $(1 \leq i \leq n)$ における出現回数を C_i とすれば，その出現回数の標準偏差 S_k は (1) 式によって求められる．

$$S_k = \sqrt{\frac{\sum_{i=1}^{n} \left(C_i - \overline{C}\right)^2}{n-1}} \qquad (1)$$

この標準偏差 S_k のままでは，出現回数の和 $(\sum_{i=1}^{n} C_i)$ が影響を与えるため，この出現回数の和を用いて正規化した値を出現頻度指標値 (F) と定義する．

$$F = \frac{S_k}{\sum_{i=1}^{n} C_i} \qquad (2)$$

(2) 式により，F の値域は $0 \leq F \leq 1/\sqrt{n}$ となる．本削除手法では，この出現頻度指標値（F）が出現頻度閾値（T）に満たないキーワードを認識語彙候補から削除することで認識語彙数削減を狙う．

2.2.2 複合名詞構成名詞の削減

適切なキーワードを抽出するために，ユーザの発声する可能性の高いキーワードとして “複合名詞” という概念を取り入れた．これはまた，複合名詞を構成する個々の名詞については発声される可能性が低いことを意味する．そこで，複合名詞（Ex. イラク復興支援会議）が抽出された場合，その複合名詞を構成する名詞（Ex. イラク, 復興, 支援, 会議）は認識語彙として抽出しないこととする．なお，別の場所で単独でその名詞（Ex. イラク）が現れた際には，その名詞（Ex. イラク）は認識語彙として登録される．

2.2.3 数名詞の削減

モバイル環境では，手元に参照するデータが存在することは少ないと考えられる．また逆に，数字を含むキーワードを知りたいために検索するのであって，数字を含むキーワードで検索することは少ないとも考えられる．そこで，正確な数字を認識語彙として設定しても実行上意味がないと考え，数名詞を始め，数字を含む名詞については削除することとした．

2.2.4 短単語の削減

「意味の乏しい短いキーワードを発声する可能性は低いだろう」という，複合名詞と対となる仮定に基づくのが本手法である．本稿では 2 モーラ以下のキーワードについては，発声可能性が低いとして認識語彙から削除することとした．

3 実験評価

本章では，前章で提案した “複合名詞” の導入によるキーワード抽出手法と，4 つの認識語彙削除手法を併用した認識語彙抽出手法に対して，被験者の発声キーワードによる有効性評価方法およびその評価結果について説明する．

3.1 評価方法

被験者に対して「最近のニュースを検索をするためにキーワードを発声して下さい」という要求を行い，その結果を評価元データとした．延べ 132 名の被験者から 417 キーワードを収集した．なお，本研究の目的を考慮して，被験者にはキーワードを記述するのではなく発声するよう指示し，それを録音したものからテキストに起こしたものを評価元データとして利用した．

今回は提案手法の評価を主眼とするために音声認識率を 100%として，発声されたキーワードと完全一致するキーワードが認識語彙の中にあったかどうかを評価した．なお，音声認識の特性上，キーワードの「ヨミ」が認識対象となるため，被験者発声から起こしたヨミと，キーワード抽出の際に形態素解析機である茶筌が出力するヨミとが完全一致する場合に，「一致した」と判断する．従って「松井稼頭央」などの難読キーワードについては，キーワードとして抽出されていても茶筌が出力するヨミと一致しないため，本評価では「一致せず」と判断される．本課題については，頻繁に出現するキーワードのヨミを適宜メンテナンスすることで，比較的容易に解決可能であると考えられることから本稿では言及しない．

本稿では評価元データと一致するキーワードを検索する対象記事として，記事登録時間が分単位で表示されている毎日新聞社の Web サイト [8] の記事を用いた．被験者がキーワードを発声した時間から，6,12,18,24,30,36,42,48 時間前までさかのぼった期間に登録された記事を検索対象として評価した．すなわち「24 時間」であれば，被験者発声時間から 24 時間前までに登録された記事を検索対象として評価した．

被験者の発声データ収集は 2004 年 3 月 10 日から 3 月 12 日の 3 日間にわたって実施し，カテゴリ毎の出現回数の集積は 2004 年 3 月 7 日の記事から，各発声データの発声タイミングの前までの記事における各キーワードの出現回数を用いた．また，カテゴリは毎日新聞社の Web サイトの分類に従って，「社会」「経済」「スポーツ」「国際」「政治」「人事」の 6 種とした．従って，出現頻度指標値 F の値域は $0 \leq F \leq 0.408$ となる．

3.2 評価尺度・比較対象

本稿の目的とする語彙数削減に対する提案手法の効果を評価するため，認識語彙数を評価尺度とする．また，発声キーワードと一致する認識語彙が存在するかどうかという Hit 率についても評価尺度とする．

提案手法によって抽出された認識語彙による Hit 率および認識語彙数と以下のものをそれぞれ比較評価する．

(1) 検索キーワードに一致する記事が存在するか，という観点から，インタネット上の検索エンジンなどで多く利用されるテキストベースの検索結果を Hit 率の比較対象とする．本稿ではシソーラスなどは用いずに，単純に見出しおよび本文の部分一致全文検索を行った結果と比較した．

(2)「その記事内容をよく表しているのが見出しである」という仮定から，見出しから抽出されたキーワードが検討されることがある．そこで，見出しから抽出した名詞による認識語彙を用いて，特に Hit 率について比較評価する．

(3) 本提案手法のキーワード抽出手法の有効性を評価するため，見出しおよび本文から抽出した名詞を認識語彙とした場合の Hit 率および認識語彙数について比較評価する．

(4) 既存手法である TF-IDF 法による効果とを比較評価する．提案手法とほぼ同数の語彙数まで TF-IDF 法にて語彙削減を行った際の Hit 率について提案手法と比較する．

(5) 本提案手法の認識語彙数低減手法の有効性を評価するため，見出しおよび本文から抽出した複合

表 2: 検索対象時間毎の平均検索対象記事件数

検索対象時間	平均検索対象記事件数
6	43.53
12	68.62
18	104.52
24	186.40
30	228.31
36	250.47
42	282.42
48	362.97

名詞から数名詞，短名詞を削除しない認識語彙の場合の認識語彙数について比較評価する．また，認識語彙削減によって Hit 率がどの程度減少するかについても評価する．

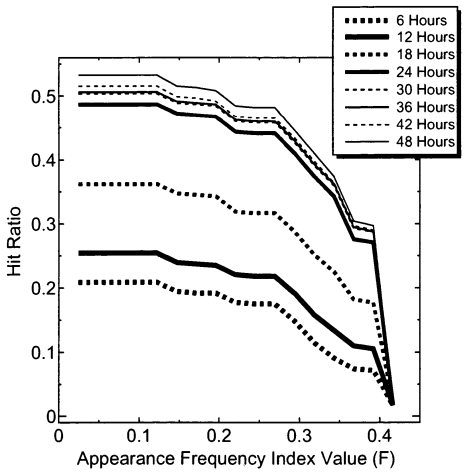

図 1: 出現頻度指標値による Hit 率の推移

3.3 評価結果

提案手法を評価した結果について以下に説明する．評価対象期間には計 5371 件の記事が存在した．また，検索対象時間ごとの平均検索対象記事件数を**表 2** に示す．

3.3.1 出現頻度閾値（T）の設定

出現頻度指標値（F）を用いた提案手法では，出現頻度閾値（T）の設定が重要な意味を持つ．そこで出現頻度指標値 (F) による Hit 率の違いの推移を調査し，妥当な出現頻度閾値（T）を求めることとする．出現頻度指標値による Hit 率の推移を**図 1** に示す．

横軸に出現頻度指標値（F），縦軸に Hit 率を取り，各検索期間における Hit 率の推移を示した．図 1 によれば，$F = 0.12$ までは Hit 率の落ち込みが観測されず，$F > 0.12$ において Hit 率が悪化している．そこで，以降においては，出現頻度閾値 $T = 0.12$ と設定し，比較評価を行う．

3.3.2 認識語彙数

音声認識率とトレードオフの関係にあることから，可能な限り少数であることが望ましい認識語彙数について比較評価結果を**図 2** に示す．横軸に検索対象とした期間，縦軸に認識語彙数を表す．上記の通り提案手法による語彙削減においては，出現頻度閾値は $T = 0.12$ として計算した．

提案した認識語彙数削減手法を用いない複合名詞に対して，35%程度の認識語彙数削減の効果があることが示されている．

3.3.3 Hit 率

各種キーワード抽出手法による Hit 率の評価結果を**図 3** に示す．横軸に検索対象とした期間，縦軸に Hit 率を表す．

図 3 によれば，どのような手法を用いたとしても 6 時間前から 24 時間前にかけて Hit 率が急激に上昇し，その後横ばいになっている．従って，一般的に「最近のニュース」といった場合に約 1 日前までのニュースをヒトは思い浮かべることが示されていると考えられる．そこで，最近のニュース検索をするサービスを考える際には，発声時から 24 時間前までのニュース記事を検索対象とすることで十分であるとも考えられる．

一方，提案手法による Hit 率向上に関しては，見出し文や見出し文＋本文から抽出した名詞を認識語彙とした場合と比較して 50%から 190%，平均 117%の Hit 率向上が確認されており，利用者は意味あるキーワードとしての複合名詞を発声する可能性が高い，という仮定を裏付ける結果となった．

一方，テキストベースの全文検索と比較すると，14%から 32%，平均 20%程度の Hit 率下落にとどまった．また，認識語彙数削減手法を用いない複合名詞と比較した場合には，2%程度の Hit 率下落に過

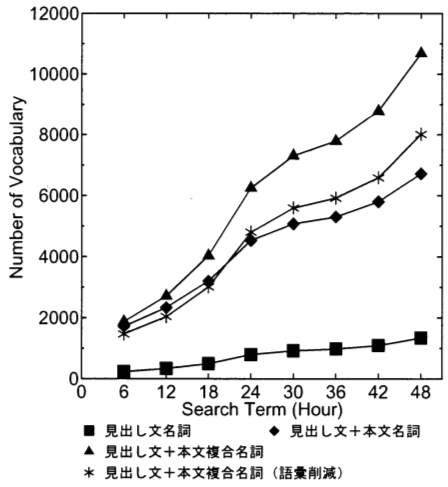

図 2: 各種抽出方法による認識語彙数の評価結果

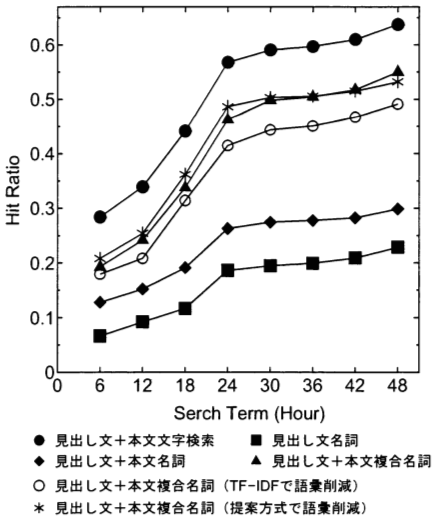

図 3: 各種抽出方法による Hit 率の評価結果

ぎず，数名詞および短い単語，多くの記事に出現するキーワードは発声されにくいという仮定が検証されたと考えられる．さらに従来手法である TF-IDF 法を用いて提案手法と同等の語彙数まで認識語彙を削減した場合の Hit 率と比較すると，平均 15%程度の向上が見込まれる．

また，Hit 率の絶対値について，テキストベースの全文検索であっても 6 割程度の Hit 率である理由として，形態素解析のヨミ付与が失敗した難読キーワードや元々ニュース記事にも存在しないキーワードが発声された場合が比較的多かったことが上げられる．サービスとしては，キーワード例を示すインタフェースを採用することでユーザを誘導し，本課題を回避できる可能性が高いと考えられる．

4 まとめ

本稿では，モバイル環境をターゲットとして，ニュース記事を音声認識にて検索する際の音声認識語彙をニュース記事から抽出する手法について提案し，被験者評価を通して提案手法の有効性を示した．

提案手法は，連続した名詞からなる “複合名詞” に着目して認識語彙を増やすと共に，複合名詞中の部分名詞，数字を含む名詞，2 モーラ以下の短い名詞，様々なカテゴリのニュース記事に出現する名詞を削減することで，認識率とトレードオフの関係にある認識語彙数の削減を行うものである．

被験者発声による計 417 キーワードに対する評価結果において，複合名詞導入による Hit 率の向上と各種認識語彙数削減手法による認識語彙低減の効果が従来手法である TF-IDF 法を上回ることが確認された．

参考文献

[1] 増井 俊之, “携帯端末のテキスト入力手法”, ヒューマンインタフェース学会誌, Vol.4, No.3, 2002.

[2] 荒金 陽助, 久保田 浩司, “自動車内における情報入出力インタフェースの現状と課題”, 情報処理学会マルチメディア, 分散, 協調とモバイル（DICOMO2003）シンポジウム, 7F, pp.661–664, 2003.

[3] 岩崎 知弘, 難波 利行, 石川 泰, “カーナビゲーション用音声インタフェース技術”, 自動車技術, Vol.57, No.2, pp.65–70, 2003.

[4] NTT Communications, “V ポータル Web ページ”, http://www.ntt.com/v-portal/.

[5] K.S. Jones, “A statistical interpretation of term specificity and its application in retrieval”, J. Documentation, vol.28, No.1, pp.11–20, 1972.

[6] 松田 透, 小川 泰嗣, “統計的確率に基づくキーワード重要度算出モデル”, 情処研報, Vol.96, No.87 (NL–115), pp.123–128, 1996.

[7] 奈良先端科学技術大学院大学 情報科学研究科 自然言語処理学講座, “形態素解析システム茶筌 FrontPage”,http://chasen.aist-nara.ac.jp/hiki/ChaSen/

[8] 毎日新聞社, “Mainichi Interactive”, http://www.mainichi-msn.co.jp/