報告

03

ウィキペディアデータを利用した意味的キーワード抽出手法

苗村昌秀　　山内結子

Semantic Keyword Extraction Using Wikipedia Data

Masahide NAEMURA and Yuko YAMANOUCHI

要　約

放送通信融合に向けた取り組みとして，視聴者の視聴状況を自動的に取得・分析して興味を推定し，それに応じて番組の関連情報を効果的に提示する技術の研究を進めている。番組の関連情報の取得は，基本的に，字幕データから抽出したキーワード情報を基にしているが，効果的な関連情報の提示を実現するためには，視聴者の興味に関係する意味までも含んだキーワードの抽出が重要となってくる。そこで，ウィキペディアのデータを利用して，抽出対象のキーワードとウィキペディアのページ見出し語との関係を構造化したコーパスを構築して，意味的な（興味に関係する意味まで含む）キーワードの抽出を行う手法を開発した。従来手法では，字幕データでよく出現する複合語の抽出や，人物のあだ名に代表される同じ意味の名寄せ処理などが課題となっていたが，提案手法ではこれらの課題を解決し，抽出するキーワードへの効果的な意味付けを可能とした。評価実験を実施し，実際のテレビ番組の字幕データから提案手法により抽出したキーワードと，視聴者が興味を持ったキーワードとの一致度を評価したところ良好な結果が得られた。本稿では，提案手法の詳細と実験結果について報告する。

ABSTRACT

We are researching technologies for the automatic acquisition and analysis of viewing status and the estimation of viewers' interests in a domestic TV viewing environment, with a view to developing new services that provide program related information according to viewers' preferences. In order to infer viewers' interest targets, we developed a method to selectively extract keywords from closed caption data that belong to particular classes derived from Wikipedia category information. This method enabled selective extraction of keywords identifying the names of people and places, which are generally promising candidates for viewer interest. The estimation experiment was made to verify the effectiveness of the proposed method, in which the matching degree between the extracted keywords and viewers interested ones was checked.

1．はじめに

テレビの視聴環境の多様化に伴い，スマートフォンやタブレット端末からインターネットの情報にアクセスしながらテレビを視聴するスタイルが一般的によく行われるようになってきている[1)]。しかし，従来のテレビ視聴環境では，例えば番組の登場人物の情報を調べる場合に，視聴者自身がネット端末にその情報を入力する必要があるなどの手間がかかり，放送と通信の間で情報をシームレスにつなぐ仕組みは不十分であった。そこで，放送と通信の融合を促進する仕組みとしてハイブリッドキャストが提案され[2)]，すでにその機能を搭載したテレビ受信機も販売されており，放送と通信の垣根を低くするインフラ的な環境は整備されつつある。今後の普及の課題は，その環境で動作するアプリケーションがどのくらいユーザーにとって使いやすいテレビインターフェースを提供できるかといった問題に移ってきている。

そこで我々は，テレビを視聴しているときに簡単にそのテレビ番組に関連したインターネット上の情報にアクセスが可能なテレビインターフェースを提案している[3)]。将来的には，このインターフェースを発展させ，視聴者の番組への興味内容を蓄積・解析して個人の嗜好（しこう）情報を取り出し，個人の好みに適した情報を提供するテレビの個人適応を視野に入れて研究を進めている。このような技術を実現するためには，個人の視聴行動から番組に対する興味内容を推測できる仕組みが必要となってくる。視聴者の視聴行動の分析は，リモコンの操作履歴や，顔の向きや表情認識などの画像認識技術から推定することができる[4)]。また，視聴者が興味を持った番組内容の推定は，番組メタ情報（番組の属性を表す情報）を取り出して，そのときの視聴行動と対応付けることにより推定できる。

番組メタ情報としては，現状では一般に利用可能なものとして，番組の字幕データから抽出したキーワード（Key Word：以下，KW）が広く用いられている。このとき，抽出したKWから関連情報を提示するといったKWの意味に立脚した処理をスムーズに行うためには，いかにそのKWの意味を表す「エンティティ」*1まで含めた意味的なKW抽出が行えるかが重要となってくる。

本稿では，上記の課題のうち後者の，番組の字幕データから意味的なKWを抽出する課題に対する解決手法を提案する。提案手法は，オフライン処理とオンライン処理から構成される。

オフライン処理では，ウィキペディアの説明文を解析することにより，ウィキペディアの見出し語に加えて，関連語句までコーパス*2を独自に拡張し，抽出するKWの範囲を広げている。

オンライン処理では，番組ごとにEPG（Electronic Program Guide）データ中の番組説明文から，エンティティが特定できるKWを基に，番組単位のローカルなコーパス（ローカルコーパス）を構築する。そして，前述の拡張したコーパスとローカルコーパスを関係付けて併用することにより，字幕データからの意味的なKW抽出を実現している。

提案手法の有効性を確認するために，テレビ視聴実験を実施し，視聴者の選択した興味対象のKWが，提案手法により，どれだけ字幕データから自動抽出できているかを調べた。

本稿では，ウィキペディアを利用した字幕データからの意味的なKW抽出手法の詳細と，その評価結果について報告する。

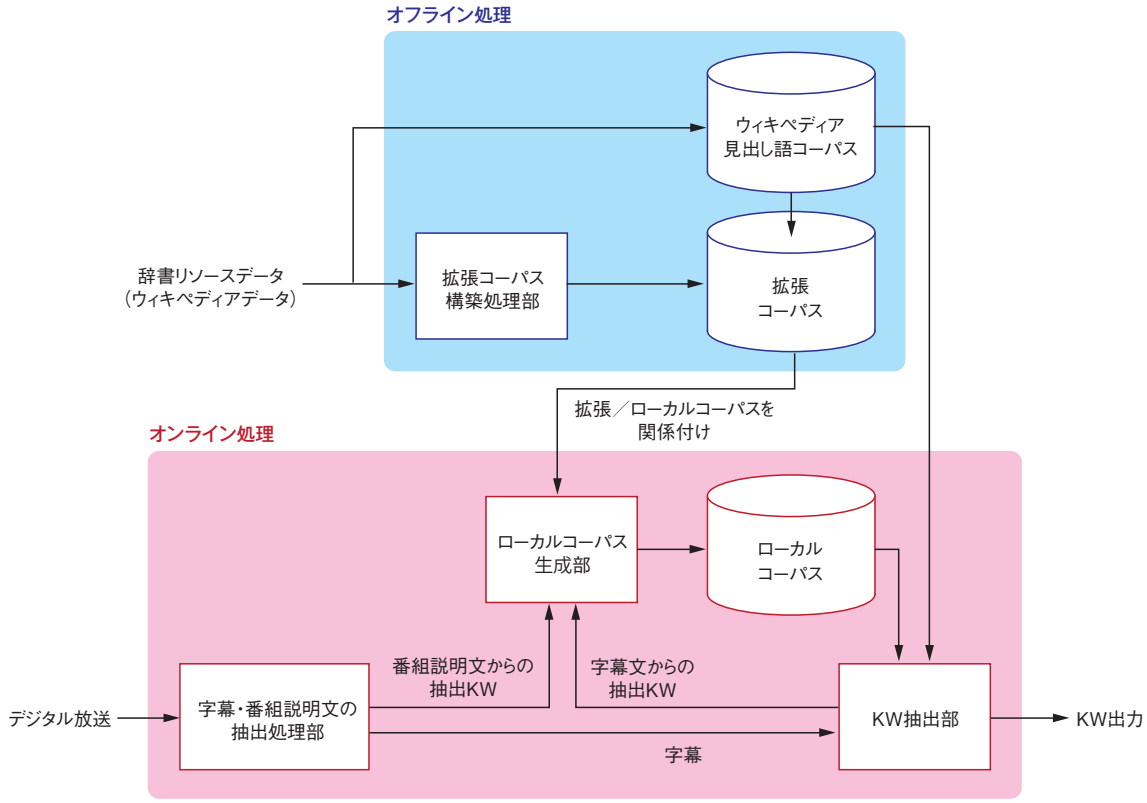
2．これまでの関連研究

字幕データなどのテキスト文からKWを抽出する処理は，番組や文書の検索，類似番組や関連情報の推薦処理の過程においてよく行われている。通常は，テキスト文をMeCab[5)]などの形態素解析*3処理で品詞分類して名詞句を取り出すことを基本としており，取り出したKWのエンティティを特定する処理は，別に用意されたコーパスデータとの照合処理や，対象文書に明示的に定められたルールに従って行っている。例えば，字幕データから抽出したKWを映像再生装置の再生時に利用して，簡単に再生内容が分かるように，再生コンテンツのナビゲーションを行っている先行例がある[6)]。この例におけるKW抽出処理は，字幕データに形態素解析を行った後，括弧やクオーテーションなどの記号で囲まれている名詞部分を重要語として抽出する処理と，その名詞の意味があらかじめ作成しておいたジャンルデータベースに登録されているかを調べて抽出する処理から構成されている。このように従来のKW抽出技術では，形態素解析処理が基本となっており，複合語，外来語，固有表現，口語文への対応などが課題となっていた。一方，提案するKW抽出手法は，後述のように，膨大なKW集合を有するコーパスの要素KWごとの照合処理を基本としているため，上記の形態素解析による課題の影響が少なく，重

＊1　一般には，「対象物の本来の意味を表す概念」を指す。本稿では，人や物を特定するための「語義が一意に定まる概念」として用いる。

＊2　自然言語処理に用いるために，自然言語の文章の言語的な情報を構造化し，大規模に集積したもの。

＊3　文章を意味のある最小単位の単語に分割する技術。



1図　提案手法の全体システム構成

要なKWを高い確率で抜き出すことが可能である。

抽出したKWとエンティティの関連付けは，KWに対して意味を関連付けてコーパスを構築し，KW抽出時にその関連性を利用するのが一般的である。このようなコーパスを構築するために，ウィキペディアのようなオープンリソースを効果的に活用することで解決を図っている研究例が多い。その代表的な例として，参考文献[7)8)]の研究がある。これらの研究に触発されて，より高度な知識処理で文書内に出現するKWのエンティティをウィキペディアの見出し語として特定する研究が多く見受けられる[9)~11)]。我々の提案手法も，基本的には，ウィキペディアの見出し語を対象とするものであるが，ウィキペディアの見出し語に対して，その記事の本文構造を利用して関連する語句までコーパスの登録KWを拡張させている点が，従来手法と異なっている。

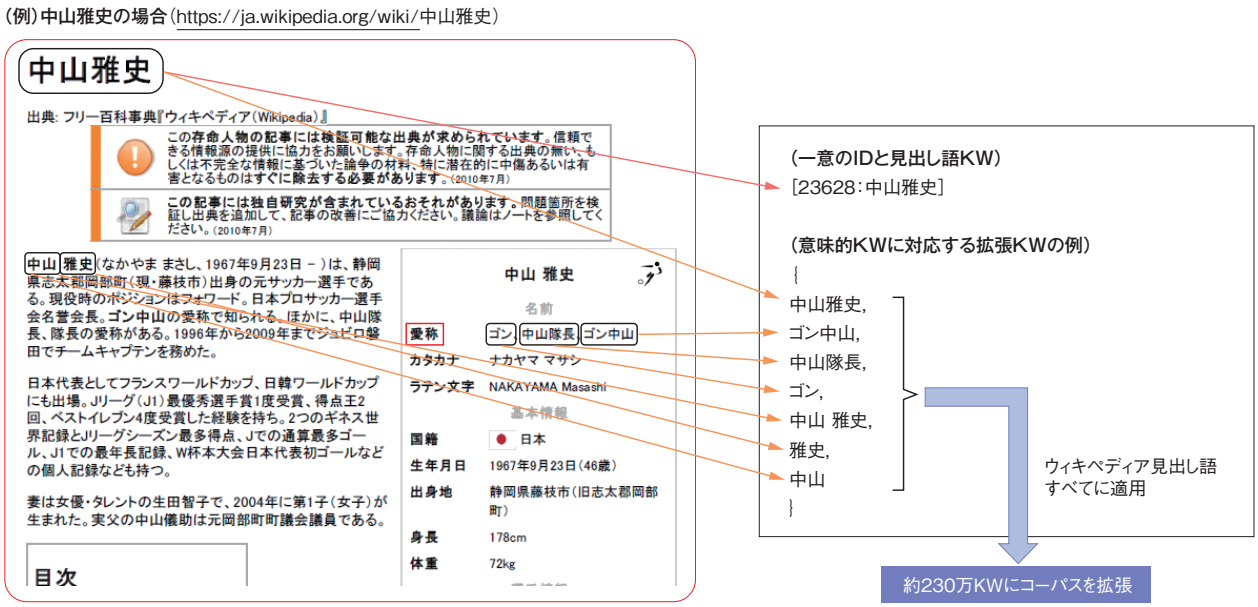
3．字幕データからのKW抽出手法の提案

3．1　全体概要

提案手法における意味的なKW抽出は，字幕データから抽出した，語義の曖昧性を有するKWを，ウィキペディアの見出し語に関連付けることで実現している。1図に提案手法の全体システム構成を示す。1図に示すように，処理の構成は，事前に必要なコーパスを構築しておくオフライン処理部と，リアルタイムで入力データを処理するオンライン処理部に分かれる。

事前に必要なコーパスは，ウィキペディア見出し語コーパスと，ウィキペディア見出し語に関係するKWを集めた拡張コーパスから構成される。オフライン処理部では，ウィキペディアの本文を解析してこの拡張コーパスを構築する。

オンライン処理部では，処理中の番組の番組説明文から抽出したKWを拡張コーパス内のKWと関係付けて，番組に固有のローカルコーパスを構築する。番組説明文は，デジタル放送に多重されているEPGデータに含まれる番組内容を記述したテキストで，その番組の出演者や地名などの情報がウィキペディア見出し語と一致する形式で記載されていることが多い。そのため，番組説明文を基に構築したローカルコーパス中のKWを活用することにより，字幕データから抽出したKWが名前の言いかえや短縮形などである場合でも，そのエンティティを特定することができる。オンライン処理部の入力は，デジタル放送であり，デジタル放送に多重されている字幕データとEPGデータ内の番組説明文の部分を抽出する。デジタル放送から抽出された字幕データは，KW抽出部に導かれ，そこでオフラインで構築したウィキペディア見出し語コーパスと番組ごとのローカルコーパスのKWとの照合処理を行って，字幕データに含まれるKWが検



2図　ウィキペディアページからの拡張KWの取得例

出される。KW間の照合処理は，3．4節で述べるように，語長の長いKWから段階的に照合していく処理である。これは，語長の長い単語ほど信頼性が高い，という仮定に基づき，形態素解析による検出の揺らぎを防ぐ目的で行っている。

3．2　ウィキペディアからの拡張コーパスの構築

日本語のウィキペディアでは，140万語以上のページがページの見出し語を基に登録されている。また，新しいページも日々生成・更新され，英語版では内容についても信頼性が高いとの評価を受けている[12]。データ自体はXML（Extensible Markup Language）フォーマットで構造化されており，それぞれのページの見出し語の別表現を表すリダイレクト情報や，見出し語の属するカテゴリー分野を表すカテゴリー見出し語などの区別も識別子などを用いて管理されている。また，見出し語の種類によっては，その内容を簡単に説明するインフォボックスなどが付加されているものもある。**2図**に，ウィキペディアのページの一例と拡張KWの取得例を示す。2図では，見出し語に当たるのが“中山雅史”であり，“愛称”，“カタカナ”などが記された領域がインフォボックスである。この例では，ページの本文やインフォボックスを解析して，“ゴン中山”，“中山隊長”などの拡張KWを取得している。

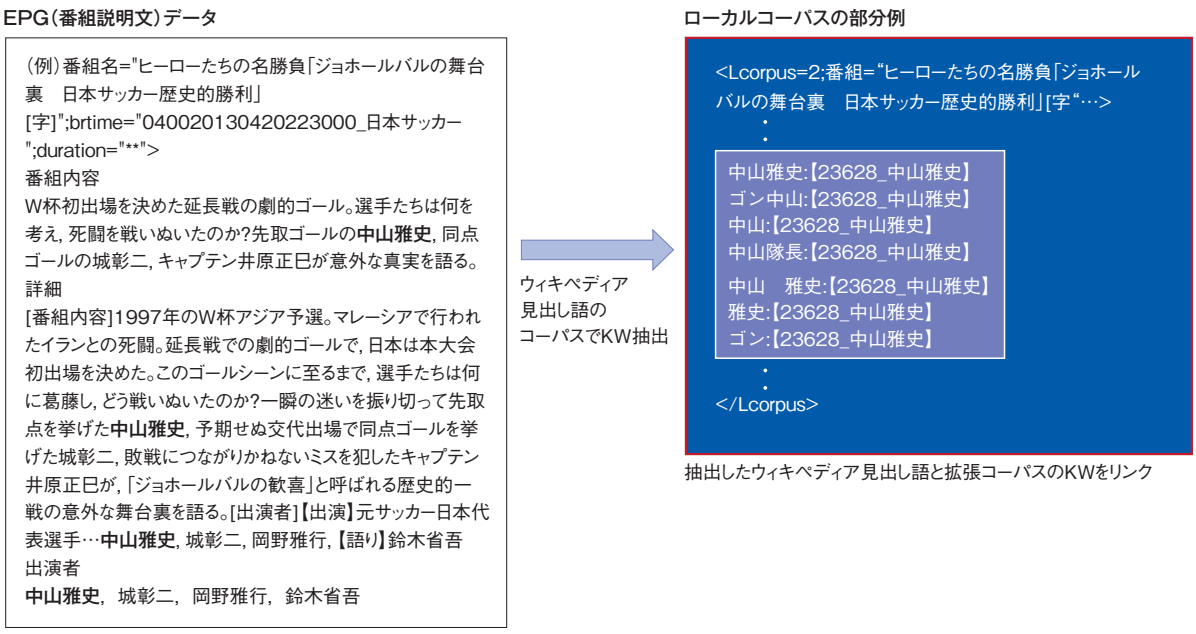
このように，構造化されたウィキペディアの特徴を活用することにより，ウィキペディアの見出し語に関係するKWを収集して拡張コーパスを構築した。放送番組の字幕データは，一般的な文書とは異なり，口語的な表現が多く，対象となるKWの言いかえが頻繁に行われる。特に，興味の対象となる傾向が高い人物の表現は，フルネームで言及されることが少なく，姓，名のみで呼ばれることが多い。また，愛称や通名も多く使われる。例えば，2図の“中山雅史”は，ニックネームとして“ゴン”と呼ばれることが多い。そのため，字幕データからエンティティまで含めてKW抽出するためには，“ゴン”という語を抜き出し，なおかつ，それが“中山雅史”であることまでリンク付けることが必要となってくる。このため提案手法では，単純にウィキペディアの見出し語だけでKW抽出用の辞書コーパスを構築するのではなく，字幕データにおけるKWの言いかえに対応できるように，分離されたKWやニックネームなどへのリンク付けを行った拡張KWを加えて，拡張コーパスを構築した。拡張のための処理は，以下の規則に従い，ウィキペディアの内容を解析して行った。

（1）リダイレクト情報の活用

ウィキペディアのデータ構造として元来から存在するリダイレクト情報を活用してコーパスを拡張する。2図の場合，見出し語の“中山雅史”に対して，別表現である“ゴン中山”が“中山雅史”にリンクする形式でリダイレクト情報として管理されている。

（2）人物の姓名の分離

ウィキペディアでは，最初の序文で簡単にその見出し語に関する説明が行われている。このとき，対象の見出し語が人物名である場合，漢字表記の場合は，姓，名の間に空白，カタカナ表記の場合は“・”記号で区別する，といった暗黙的な規則がある。この規則を利用して，見出し語が人物である場合は，姓と名の部分を取り出し，



3図　ローカルコーパスの構築例

元の見出し語と関連付けてコーパス化する。なお，見出し語が人物に関するものであるかどうかの判別は，見出し語が属するカテゴリー情報を解析することにより可能である*4。

（3）愛称などの言いかえ語のリンク付け

見出し語によっては2図のようにインフォボックスが存在し，その中で愛称やニックネームなどの言いかえ語が記述されている場合がある。そこで，インフォボックス内の言いかえ語を抽出して見出し語に関連付けてコーパス化する。

2図の例は，“中山雅史”のウィキペディアページの抜粋と上記（1）～（3）の規則に従って構築した拡張コーパスKW要素を表す。すべてのウィキペディア見出し語に同様の処理を施し，約230万語のKWから成る拡張コーパスを構築した。ただし，拡張コーパスに登録されているKWはウィキペディアの見出し語とは異なり，複数の見出し語と関連付けられているため，拡張コーパスでKW抽出を行った場合は，KWの意味づけが一意とはならない（例えば,“中山”で抽出した場合, 複数の“中山”が存在する)。そこで，提案手法では，次節で述べるように，番組単位でローカルなコーパスを構築して，意味が一意に定まるウィキペディア見出し語コーパスに関連付けることにより，この問題の解決を図っている。

3.3　ローカルコーパスの生成手順

3図にローカルコーパスの構築例を見出し語“中山雅史”の場合について示す。ローカルコーパスは，番組に依存したコーパスであり，システムの動作としては，番組が終了したり，チャンネルが切り替わったりすると，そのつど，現在視聴している番組用のローカルコーパスが処理対象となるように切り替わる。3図の例では，番組の開始時に，デジタル放送のEPGから抽出した番組説明文の中の主要なKWを，ウィキペディアの見出し語だけのコーパスから，次節で述べる段階的なKW照合処理を用いて抽出する。

番組説明文には，出演者情報などが記載されており，ウィキペディアの見出し語と完全に一致する傾向が比較的高い。ウィキペディアの見出し語は一意に定まり，KWのエンティティと同等と見なすことができる。この性質を活用して, 抽出したウィキペディアの見出し語と，複数の拡張コーパスに登録されている拡張KWとを関係付けて，ウィキペディア見出し語と拡張KW群の同一関係を記述した形式でローカルコーパスに登録する。3図の例では，“中山雅史”という見出し語に対して，「“中山雅史”,“ゴン中山”,“中山”,“中山隊長”,“中山 雅史”,“雅史”,“ゴン”」という拡張KWが関係付けられてローカルコーパスに登録される。以後，このローカルコーパス内の拡張KWを抽出した場合は，それを“中山雅史”と見なすことにより，拡張コーパスの語義の曖昧性を解消することができる。

また，番組の字幕データからKW抽出を行って，そのKWが一意にウィキペディアの見出し語に関連付けられ，かつ，そのKWがローカルコーパスに登録されていない場合は，そのKWをローカルコーパスに新規登録し

＊4　人物KWは，そのカテゴリー情報に“年生”，“年没”，“存命人物”の文字列があるかどうかで判別できる。

て，ローカルコーパスの範囲を広げる。さらに，放送番組ごとにローカルコーパスを構築・管理することにより，番組の終了やチャンネル切り替えで番組切り替えが生じた場合でも，新しい番組に対応したローカルコーパスに切り替えることにより対応が可能である。

3．4　段階的なKW照合処理

段階的なKW照合処理を**4図**の手順に従って説明する。4図に示すように，入力は字幕データであり，事前にローカルコーパスとウィキペディア見出し語コーパスをその構成要素が語長順に並べられた状態で準備しておく。段階的なKW抽出の条件として，語長のしきい値Kを設定する。

字幕データからのKW抽出は，次のように行われる。抽出対象KW（コーパスの構成要素）の語長がK以上の場合は，ローカルコーパスおよびウィキペディア見出し語コーパスの構成要素と字幕データとの間でKW照合が行われる。一方，抽出対象KWの語長がK未満の場合は，まず字幕データに対して形態素解析処理を施してから，名詞部分と判定された部分と，ローカルコーパス，ウィキペディア見出し語コーパスの順番でその構成要素とのKW照合を行って，語長がKより小さいKWを抽出する。

対象とする字幕データの処理が終了すると，デジタル放送から読み込んだ次の字幕データについて同様の処理を行う。KW照合処理では，ローカルコーパスの構成要素を，ウィキペディア見出し語コーパスの構成要素より先に照合し，同一と判定されたKWの部分を元の字幕データから省いて，処理を続ける。これは，KW抽出の重複を避けるための処理である。また，ウィキペディア見出し語コーパスの構成要素と同一と判定されたKWについては，3．3節の処理を施してローカルコーパスに登録する。

このような過程で抽出したKWは，単にKWのみが抽出されているのではなく，そのエンティティであるウィキペディアの見出し語に関連付けられているため，視聴者へのサービスとして，抽出KWを用いた情報提示を行うときに，カテゴリー分けした提示や対象KWのウィキペディア情報の紹介など，処理結果の効果的な見せ方が容易に実現できる。

4．番組視聴実験による提案手法の評価結果

4．1　番組視聴実験

提案手法の有効性を確かめるために，実際の番組の字幕データからKWを抽出して比較評価を行った。比較に用いたKW抽出方法を以下に列挙する。

入力：字幕文
事前準備：ウィキペディア見出し語コーパス
　　拡張コーパス
　　番組ごとのローカルコーパス
抽出条件：語長しきい値Kを設定
出力：ウィキペディアの見出し語へのリンク付きKW

段階的KW抽出
　ステップ1：KWマッチング処理
　　ケース1：抽出対象KWの語長がK以上の場合：
　　字幕文と
　　・ローカルコーパスのKWとのマッチング処理
　　・ウィキペディア見出し語コーパスのKWとのマッチング処理
　　マッチングKWを出力
　　字幕文からマッチングKW部分を削除
　　ケース2：抽出対象KWの語長がK未満の場合：
　　形態素解析処理後，名詞部分を抽出し，その名詞部分と
　　・ローカルコーパスのKWとのマッチング処理
　　・ウィキペディア見出し語コーパスのKWとのマッチング処理
　　マッチングKWを出力
　　字幕文からマッチングKW部分を削除
　ステップ2：ローカルコーパスへの登録処理
　　ウィキペディア見出し語コーパスのKWとマッチングしたKWから
　　作成した拡張KWをローカルコーパスに追加登録

4図　段階的KW抽出の手順

Cmp1：形態素解析後の名詞部分をKWとして抽出
Cmp2：番組ごとのローカルコーパスなしで段階的にKWを抽出
Ref：主観的に手動で字幕データからKWを抽出
提案手法：番組ごとのローカルコーパスを併用して段階的にKWを抽出

また，提案手法を用いて抽出したKWが視聴者の興味対象のKWをどれだけ含んでいるかを調べるために，複数の実験参加者による番組視聴実験を実施した。この番組視聴実験では，番組視聴後に実験参加者にその番組内で興味をもったKW（以下，興味KW）を選択してもらい，字幕データからのKW抽出結果が，その興味KWをどのくらいカバーしているかを評価した。番組視聴実験の主な条件を以下に列挙する。

番組数：87番組（延べ20時間）
実験参加者：8名
選択してもらった興味KWの数：
　番組ごとに
　（a）上位3つの興味の高いKW
　（b）任意の数の興味のあるKW

また，番組ごとの興味カバー率CRateは次式により算出した。

$$CRate = (ExtKW \cap CuriKW) / CuriKW \quad \text{——（1）}$$

1表　興味カバー率CRate（%）

	上位3つの興味の高いKWを選択してもらった場合	任意の数の興味のあるKWを選択してもらった場合
Cmp1	72.7	75.3
Cmp2	77.6	79.6
Ref	81.1	81.9
提案手法	83.7	85.0

ここで，ExtKWは提案手法により字幕データから抽出したKWの数，CuriKWは実験参加者が自己申告した興味KWの数である。

4.2　評価結果

提案手法で抽出したKWの例を，元の字幕データとともに以下に示す。なお，抽出KW例においては，ウィキペディア固有の識別番号を付けている。

字幕データの例：

フォワードの先発はエースの三浦知良と22歳の城だった。

抽出KW例：

｛70246:フォワード，630294:先発，392153:エース，
2953:三浦知良，1260485:22歳，125367:城｝

この例はサッカーの日本代表を扱った番組で，選手名が姓のみやニックネームなどで呼ばれることが多く，従来手法では意味的な抽出が課題となっていた。提案手法では，番組説明文を解析して構築されたローカルコーパスに属するKW“城”に城彰二のウィキペディアの識別番号が付加されており，字幕データ中の言いかえに効果的に対応していることが分かる。

次に，提案手法により抽出したKWが手動で抽出したKWとどのくらい一致しているかを調べるために，提案手法と手動で抽出したKW間の適合率*5，再現率*6，および，f 値*7を計算したところ，P（適合率）= 0.6, R（再現率）= 0.9，F value（f 値）= 0.7の値を得た。この際，両者の一致については，完全一致でなくKW間で半分以上重なった場合も一致していると見なした。それぞれの計算結果より，提案手法の再現率は非常に高く，手動で抽出したKWを良くカバーしていることが分かる。適合率に関しては低い値となっているが，これは，提案手法で抽出したKWの方が，手動で抽出したKWより広い範囲のKWをカバーしている結果と考えられる。

さらに，提案手法で抽出したKWが，実験参加者が自己申告した興味KWをどのくらい的確にカバーしているかを調べるために，（1）式で定義した興味カバー率CRateを求めた。1表に，各比較手法（前節で述べた（a）または（b））における興味カバー率の値を示す。1表の結果より，提案手法が他のKW抽出手法よりも多くの興味KWをカバーしており，そのカバー率は8割以上に達していることが分かる。これらの興味KWを解析することにより，興味内容の推定手法の開発につながることが期待できる。

5．まとめ

本稿では，テレビ番組の字幕データからKWを抽出する手法を提案した。提案手法は，ウィキペディアを活用して，抽出したKWをウィキペディアの見出し語に関連付ける仕組みとなっている。また，ウィキペディア内の各種データを基に，コーパスの登録KWの拡張を行った。さらに，拡張したKWの曖昧性を解消するために，番組に固有のローカルコーパスを番組単位で構築して活用することにより，意味的なKW抽出を行う仕組みを実現した。実際の番組の字幕データに提案手法を適用し，評価実験を実施したところ，視聴者が興味を持つ可能性の高いKWを十分にカバーしていることが示された。

本提案手法には，番組説明文に記述されていないKWに関しては，有効なローカルコーパスが構築できないといった課題がある。この解決策としては，KWを抽出しようとしている文章の文脈まで考慮して語義の曖昧性を解消する方法や，推定結果をフィードバックして抽出対象KWを絞るような高度な知識処理が必要となってくる。また，字幕データ以外の情報，たとえば，番組に関する書き込みが載っているSNSやオープンキャプション*8の文章などからの情報を統合することによって，より精度の高い意味的なKW抽出が可能になると考えられる。これらは，今後の検討課題としたい。

本稿は，電子情報通信学会技術研究報告に掲載された以下の論文を元に加筆・修正したものである。
苗村，山内：“ウィキペディアデータを利用した意味的キーワード抽出手法，”電子情報通信学会，言語理解とコミュニケーション第5回テキストマイニング・シンポジウム資料，pp.63-68（2014）

＊5　検索結果として得られた集合の中に，どれだけ検索に適合した文書を含んでいるかを表す正確性の指標。
＊6　検索対象としている文書の中で，検索結果として適合している文書（正解文書）のうちでどれだけの文書を検索できているかを表す網羅性の指標。
＊7　適合率と再現率の調和平均。
＊8　映像中にスーパーインポーズされる文字（いわゆるテロップスーパー）。

参考文献

1）小島，執行：“テレビとインターネット：番組関連の同時利用の実態を探る，”放送研究と調査，2014年７月号, pp.82-100（2014）

2）A. Baba, K. Matsumura, S. Mitsuya, M. Takechi, Y. Kanatsugu, H. Hamada and H. Katoh：“Advanced Hybrid Broadcast and Broadband System for Enhanced Broadcasting Services,”NAB Broadcast Engineering Conference，pp.343-350（2011）

3）山内, 奥田, 高橋, クリピングデル, 苗村：“視聴状況に基づいた興味内容推定システムの試作，”映情学年次大, 7-6（2014）

4）M. Naemura, S. Clippingdale, M. Takahashi, M. Okuda, Y. Yamanouchi and M. Fujii：“Real-time LDCRF-based Method for Inferring TV Viewer Interest,”ACII2013，pp.485-491（2013）

5）T. Kudo：“2006 MeCab：Yet Another Part-of-speech and Morphological Analyzer,”http://mecab.sourceforge.net

6）小山，酒井，福井，上原，下森：“効率的な番組視聴を支援するための話題ラベルの生成とその評価，”情報処理学会，情報学基礎研究会資料，Vol.34，pp.17-23（2007）

7）S. Cucerzan：“Large-scale Named Entity Disambiguation Based on Wikipedia Data,”Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning，pp.708-716（2007）

8）R. Mihalcea and A. Csomai：“Wikify! Linking Documents to Encyclopedic Knowledge,”Proceedings of ACM Conference on Information and Knowledge Management (CIKM)，pp.233-241（2007）

9）E. Gabrilovich and S. Markovitch：“Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis,”Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), pp.1606-1611（2007）

10）藤井, 飯田, 徳永：“Wikipedia記事を利用した曖昧性のある表現の固有表現クラス分類，”第16回言語処理学会, 2010年３月，pp.15-18（2010）

11）B. Hachey, W. Radford, J. Nothman, M. Honniba and J. R. Curran：“Evaluating Entity Linking with Wikipedia,”Artificial Intelligence，Vol.194，pp.130-150（2013）

12）J. Giles：“Internet Encyclopedias Go Head to Head,”Nature，Vol.438，pp.900-901（2005）

苗村 昌秀（なえむら まさひで）

1984年入局。大分放送局，技術局を経て，1989年から放送技術研究所において，ハイビジョン信号処理，画像認識，ヒューマンインタラクション，データ解析技術の研究に従事。現在，放送技術研究所ハイブリッド放送システム研究部上級研究員。博士（情報学）。

山内 結子（やまのうち ゆうこ）

1988年入局。放送技術局を経て，1990年から放送技術研究所において，CG・実写映像合成処理や，コンテンツ解析および視聴状況解析処理の研究に従事。現在，放送技術研究所ハイブリッド放送システム研究部上級研究員。博士（工学）。