解説

自然言語処理技術の研究概要

田中英輝

テキストを計算機で扱う自然言語処理技術は，近年，放送局で重要性を増してきている。背景には放送局を取り巻く状況の変化がある。第１に，外国人や障害者などへのサービスの期待が高まってきたことである。第２に，放送局でインターネットを使ったコンテンツの２次利用が本格化してきたことである。第３に，インターネットで視聴者自身が意見や質問など大量の情報をテキストで発信するようになってきたことである。本稿では，これらの変化を解説するとともに，当所で行っている研究の概要を「人にやさしい放送」，「コンテンツの２次利用」，「視聴者の声の分析」に分けて紹介する。

１．はじめに

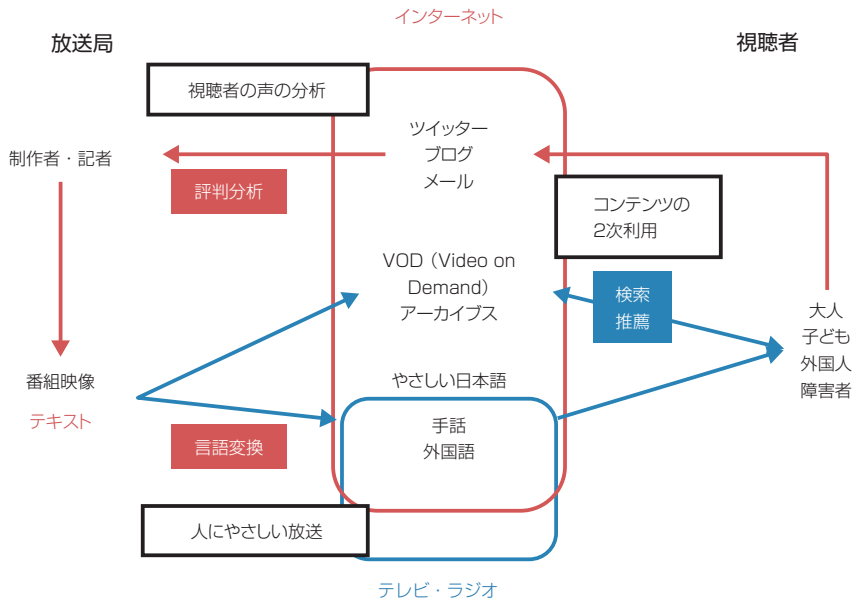
自然言語処理技術とは，入力されたひらがなの文字列を正しい漢字列に変換するカナ漢字変換システムのように，テキストを計算機で扱う技術のことをいう。この技術は，近年，放送局で重要性を増してきている。背景には放送局を取り巻く以下のような状況の変化がある。

第１の変化は人にやさしい放送への期待の高まりである。視聴者は全て同じような能力を持つ人々とは限らない。大人と子ども，目や耳に障害のある人，日本語が十分に分からない外国人など，多種多様である。それぞれの放送番組に対する受容特性（以下，特性と呼ぶ）は異なっており，特性によっては現在の放送が必ずしも理解しやすいものとはなっていない。

当所では，視聴者の特性に合わせた放送，すなわち，人にやさしい放送の研究を行っている[1)]。人にやさしい放送は障害者や外国人などから大きな期待を集めており，更なる拡大が期待されている。しかし，この放送を充実させるためには各人の特性に合わせたコンテンツを数多く制作する必要があり，コストの抑制や短時間での制作など課題は多い。自動翻訳をはじめとする自然言語処理技術はコストの低減や時間の短縮を実現する手段として大きく期待されている。

第２の変化は放送済みの番組をインターネットで配信するなど２次利用が本格化してきたことである。インターネットは視聴者単位のサービスが基本であり，実際に，インターネットでは個人が好みのコンテンツを検索し視聴している。また，放送局から個人にコンテンツを推薦するなど，多くの人を対象とした従来の放送とは違った視聴やサービスが行われている[2)]。この中心にある検索や推薦といった機能をより便利なものにするためには，自然言語処理技術が必要である。

第３の変化はTwitterなどの投稿サイトに視聴者がインターネットで膨大な情報をテキ



1図　放送と自然言語処理

ストで発信するようになってきたことである。毎日投稿されるテキストの中には，番組やNHKに対するコメントなども含まれている。膨大なテキストからNHKや番組に関連するメッセージだけを漏れなく高速に抽出し，結果を整理して表示するといった機能を実現するためには自然言語処理技術が必要である[3)]。

放送を取り巻くこのような3つの変化によって自然言語処理技術の必要性や期待感は以前より大きくなってきている。以下，当所で行っている研究を3つの変化に対応させて紹介する。

2．自然言語処理の研究概要

当所で行っている自然言語処理技術の研究を1図に示す。研究は大きく分けて「人にやさしい放送」，「コンテンツの2次利用」，「視聴者の声の分析」に分類することができる。

2.1　人にやさしい放送

人にやさしい放送を本格的に実現するためには，人の特性に合わせてコンテンツを大量に作成する必要があり，作業の効率化が課題となる。ここでは，言語を変換してコンテンツを作成する作業の効率化を目指した研究を紹介する。

（1）自動翻訳・翻訳支援

言語を変換するという作業の代表例は翻訳である。現在，NHKでは，海外放送で17か国語，国内放送で5か国語の放送を行っている。外国語の放送は基本的には日本語の原稿を翻訳して行っている。当所では，毎日大量に行われている翻訳の効率化を目指して，自動翻訳と翻訳支援の研究を行っている。

自動翻訳では日英のニュースを対象にした統計翻訳技術の研究を進めている。統計翻訳では日英の単語の翻訳確率と翻訳された文の良さを表す確率を組み合わせた値が最大になるように翻訳を行う。単語の翻訳確率と文の良さの確率をコーパスと呼ばれる大量の対訳テキストデータを使って自動的に学習する。統計翻訳は忠実に直訳されたデータで学習するほど精度が上がるが，ニュースでは翻訳時に一部分を省略することなどが多く直訳の程度は低い。このようなデータを使っても高い精度の得られる新しい翻訳知識

の学習手法を開発することが研究の目標である。

翻訳支援では日々の翻訳作業の中で作成される大量の対訳テキストを自動的に蓄積し，翻訳したい表現を検索するために使う「用例検索システム」の研究を行っている。開発したシステムは既に放送局の翻訳現場で利用されており[4]，今後，自動翻訳の機能を追加した新しいシステムを開発する予定である。

（2）やさしい日本語の放送

外国人にとっては放送を母語で視聴するのが理想である。しかし，翻訳できる言語の数に限界があり，外国人全てをカバーすることはできない。そこで，特に，国内の外国人を対象に多言語の代わりにやさしい日本語で情報を伝えようという研究が多くの研究機関で行われるようになってきている。これは，国内の外国人は一定の日本語能力を持ち，やさしい日本語で情報を伝えられると期待できるからである。

本特集号の報告「やさしい日本語ニュースの公開実験」では，ニュースをやさしい日本語で伝えるための研究の概要と2012年４月から開始したWebでのやさしい日本語ニュース「NEWS WEB EASY[5]」の公開実験[6]について報告する。

（3）手話CG翻訳

言語を変換する研究は外国人を対象にする研究だけではない。当所では，聴覚に障害のある方のために，ニュースを手話CG（Computer Graphics）へ変換する研究を行っている[7]。この研究は言語間の翻訳ではなく，言語から映像への翻訳となっている点が特徴である。

手話は，既に，一部の番組で手話通訳者によって提供されているが，番組数を更に拡大して欲しいという要望は多い。そこで，研究の最初の目標を気象ニュースに限定し，テキストから手話CGへ自動翻訳する研究を開始した[8]。しかし，手話には分かっていないことが多い。例えば，手話の言語的な調査を行う基礎データを収集するために，手話コーパスの構築を開始したが，手話動作を表す標準的な標記法がないのが現状である。

本特集号の解説「手話における言語資源の研究動向」では，手話の表記の問題，コーパス作成の動向，NHKで開発中のニュースコーパスなどを紹介する。

2.2　コンテンツの２次利用

コンテンツの２次利用が進み，現在，アーカイブスに蓄積された番組の一部はVOD（Video on Demand：ビデオオンデマンド）*1でインターネットでも提供されるようになってきている。インターネットのサービスの中心にある技術が言語を使った検索と推薦の機能であり，当所では，その研究・開発を進めている。以下，番組検索の研究を紹介する。

＊1
NHKオンデマンドなどがこれに相当する。

（1）単語間の関係の利用

EPG（Electric Program Guide：電子番組表）にある番組概要文など，番組に付随したテキスト情報を利用した番組検索の研究を行っている。現在主流の「キーワード検索法」は入力キーワードを含むEPGの番組概要文を検索して，対応する番組名を出力する手法である。しかし，この手法には入力キーワードを含む番組概要文がないときには出力が得られないという問題がある。特に，番組数が少ない場合には推薦する番組名が出力されないことが多い。そこで，キーワードの一致条件を緩和して，検索結果を増やす手法を検討している。

例えば，単語の意味的な関係を使う手法である。「地震」と「学校」で検索するときに，それらの類義語である「震災」や「教育機関」も含めて検索することで，得られる結果が増える。また，類義語ではなく原因－結果，病気－対処法，場所－名物などさまざま

な関係を持つ単語データを利用して検索を行う方法がある。このような方法を実現するためには単語の関係データが必要となるが，人手をかけて作るのでは膨大なコストと時間がかかる。そこで，多数の単語の関係をインターネットやニュースのデータベースから自動的に抽出するための研究を進めている。

（2）シーン検索

場所やゲストの紹介シーン，現地からの中継シーン，スポーツ中継の決定的な瞬間のシーンなど，番組をシーン単位で検索できると便利である。これを実現するためには，何のシーンなのかを説明した文字データがシーンごとに付いていればよい。そこで，字幕やナレーションから各シーンを説明した情報を抽出する研究を行っている。

例えば，スポーツ中継における決定的な瞬間のシーンの抽出では，アナウンサーや解説者の話した言葉を１文ごとに「試合記述文」と「解説文」に自動分類する。試合記述文とは試合状況を実況しているコメントで，解説文とは試合状況には関係していないコメントである。試合記述文のまとまりを１つのシーンとして抽出し，各シーンの重要性を統計的に判定する。この結果を利用することで，スポーツ番組のダイジェスト視聴や決定的な瞬間のシーンの検索が可能となる。このような技術をニュース番組にも応用して，現地からの中継シーンで使われる特徴的な言葉の言い回しを統計的に解析し，ニュース番組をシーンごとに管理するシステムの構築を進めている。

2.3　視聴者の声の分析

放送制作者は視聴者からの意見や感想を次の番組制作に生かす必要がある。そこで，電話・メールなどの多様な経路を通して視聴者から寄せられる意見を分析し，番組制作に生かすための情報を効率的に収集するための評判分析技術の研究を進めている。

意見が寄せられる経路としては，従来は電話やファックスが大半であったが，最近ではメールや番組Webページの意見収集フォームを使った投稿も増えている。また，意見を直接NHKに投稿するのではなく，ブログやTwitterに書くことが増えてきており，それらから情報を抽出することも重要になっている。

評判分析の基本的な処理は，与えられた文が「何に対して」「どういう意見」を述べているのかを分析することである。例えば，ある番組に対するコメント「ダイオウイカの海中シーンはこれまで見たことがなかったので感動した」から「ダイオウイカの海中シーン」に対して「感動した」という意見（反応）であることを決定することである。

意見が寄せられる経路によっては，基本的な処理の他に特別な処理が必要である。例えば，Twitterでは極端に短いコメントや砕けた表現が頻繁に使われる。また，どの番組に対するコメントなのかはっきりしないことも多い。そのため，Twitterを対象とする場合には砕けた表現への対処や番組名の推定などが必要である。

当所では，基本的な処理方法の開発と，意見が寄せられる経路に特有な問題の解決方法の両方に取り組んでいる。本特集号の報告「視聴者の意見を把握するための評判分析技術」では，評判分析の典型的なシステムと要素技術を紹介するとともに，開発した技術を放送現場に導入した例を報告する。

3．おわりに

放送局の仕事と自然言語処理技術の関係について説明し，当所で行っている研究概要を紹介した。１図に示した研究を別の視点から見ると，これらはコンテンツ制作，２次利用，制作者へのフィードバックという放送の典型的なサイクルに関係していることが分かる。

自然言語処理技術は適用範囲の広い技術である。今後も放送サービスに関連したさまざまな応用が現れると予想される。新規のサービスやアプリケーションに対応するためには，汎用性の高い言語解析などの基礎技術が重要である。今後も基礎技術とその応用技術の両方の研究を進めていきたいと考えている。

参考文献

1）今井，比留間，田中：“放送サービスへのアクセシビリティ向上に向けた取組，”ITUジャーナル，Vol.43, No.2, pp.23–27（2013）

2）住吉：“利用者の興味を拡大する，関連コンテンツ検索・推薦システム，”第23回電子情報通信学会情報伝送と信号処理ワークショップ，No.1.3, pp.19–28（2010）

3）大塚，乾，奥村：“意見分析エンジン－計算言語学と社会学の接点，”コロナ社（2007）

4）後藤，田中：“多言語翻訳用例提示システムの開発と運用，”2004年映メ学会年次大会，No.21–1（2004）

5）http://www3.nhk.or.jp/news/easy/

6）田中，美野：“やさしい日本語ニュースの公開実験サイト「NEWS WEB EASY」の評価実験，”情処研資，Vol.2012–NL–209, No.9（2012）

7）加藤，金子，井上，梅田，比留間，長嶋：“用例利用による日本語－手話CG翻訳システム，”電子情報通信学会HCGシンポジウム，I–1（2011）

8）加藤，宮崎，金子，井上，梅田，比留間，長嶋：“気象情報の日本語－手話CG翻訳，”言語処理学会年次大会，PA1–21, pp.275–278（2012）

田中英輝（たなかひでき）

1984年入局。宮崎放送局を経て，1987年から放送技術研究所およびATRにおいて自動翻訳，音声翻訳，自動要約，やさしい日本語の研究に従事。現在，放送技術研究所人間・情報科学研究部主任研究員。博士（工学）。