世界メディアブラウザ

世界中の音・映像コンテンツを日本語で視聴するシステム「世界メディアブラウザ」は，高度な音声言語処理技術によって外国語コンテンツの日本語による検索や日本語字幕付き再生を可能にします．本稿では，これを実現する私たちの世界最先端技術とマサチューセッツ工科大学の講義ビデオを用いたプロトタイプシステムを紹介します．

堀 貴明（ほり たかあき） /須藤 克仁（すどう かつひと）
塚田 元（つかだ はじめ） /中村 篤（なかむら あつし）
NTTコミュニケーション科学基礎研究所

世界メディアブラウザとは

「世界メディアブラウザ」は母国語のキーワードによって世界の動画コンテンツを検索し，そのキーワードが話されたシーンを母国語の字幕をつけて再生することを指向したシステムです．このシステムが実現すれば，世界中の価値あるコンテンツに誰もが容易にアクセスできるだけでなく，世界に向けた情報発信もより敷居の低いものになるでしょう．しかし，このようなシステムの実現には高度な音声認識技術，言語解析技術，機械翻訳技術が不可欠であり，多くの難しい課題を含んでいます．本稿では世界メディアブラウザの実現に向けた私たちの技術的チャレンジを紹介していきます．

研究の背景

近年，ブロードバンドネットワークを通じて，日本に居ながらにして，世界中の音・映像コンテンツを視聴できるようになりました．動画配信サービスによって映画やTV番組がタイムリに提供されるだけでなく，YouTube™に代表される動画共有サイトの登場により，一般のエンドユーザが自由に動画を作成・公開できる仕組みも整いつつあります．これに伴い，世界中で作成された動画コンテンツが日々インターネット上にアップロードされ，その数は急速に増加しています．

しかしながら，自分の見たい内容を含むコンテンツを見つけるのは簡単なことではありません．例えば動画共有サイトでは，各動画ファイルに付与された高々数個の関連キーワードを手掛かりに，膨大な候補の中から目的の動画を探し出さねばなりません．また，世界中のコンテンツにアクセスできる環境が整う中で，言葉の壁もまた大きな問題になっています．外国語のコンテンツは，たとえその内容に興味があっても，母国語ではない多くの人にとっては検索することも視聴することも困難です．

このような課題を解決するために，私たちは世界メディアブラウザの研究を立ち上げました[(1)]．そして，これまで研究してきた音声認識，言語解析，機械翻訳に関する世界最先端レベルの成果を結集し，そのプロトタイプシステムを構築するとともに，さらなる発展に向けた研究を進めています．

近年，米国やヨーロッパでも音・映像コンテンツの翻訳に関する研究が国家的プロジェクトとして進められ，大きな注目を集めています．私たちの研究もこれらと関連する部分はありますが，その違いはWeb上のより幅広い内容のコンテンツを対象とする点です．つまり，他のプロジェクトよりも認識や翻訳の難しい実世界のデータを対象にしています．また，海外のコンテンツを日本語に翻訳して視聴可能にすることを重視し，技術的にも日本語に適したアプローチを取っています．

システムの構成

図1は構築したプロトタイプシステムのブラウザ画面を表しています．現在は英語のコンテンツを日本語で視聴することができます．

上部に配置されたクエリ（検索キーワード）入力フォームから日本語または英語のキーワードを入力すると，そのキーワードが話されているシーンの候補一覧が右側に表示されます．キーワードが日本語，探したいコンテンツが英語であっても，コンテンツの日本語訳にそのキーワードが含まれていれば，そのシーンが候補として検索されます．シーン候補のどれか1つを選択すると，左側にその選択したシーンから動画が再生されます．ビデオを再生するときは，話された英語とその日本

語訳が字幕として動画の上下に表示されます．これらの字幕は，コンテンツの音声情報を解析することで自動的に得られた情報です．

また，動画の再生に合わせて，話された言葉に含まれる固有表現が表示されます．固有表現とは，人名，会社名といった固有名詞や，日付，金額などの数値表現を含みます．表示された固有表現はWeb検索エンジンにリンクされているため，ユーザは動画を再生しながら固有表現の意味を調べることができます．

ブラウザがこのように動作するには，あらかじめ動画ファイルを収集し，さまざまな情報を抽出しておく必要があります．これは，世界メディアブラウザのコンテンツ解析（**図2**）によって行われます．

まず，インターネットまたはローカルディスクから動画コンテンツを逐次取得します．次に各コンテンツの音声信号を音声認識することで話されている内容の文字情報を得ます．さらに音声言語解析によって連なった文字列を文の単位に分割し，固有表現を抽出します．統計的機械翻訳により，各文を目的の言語に翻訳します．音声認識，言語解析，機械翻訳の結果は，動画と同期して表示するため，時刻情報とともにアノテーションデータベースに保存されます．ブラウザは，検索サーバを介してデータベースにアクセスすることで，任意のキーワードを含むシーンを検索できます．

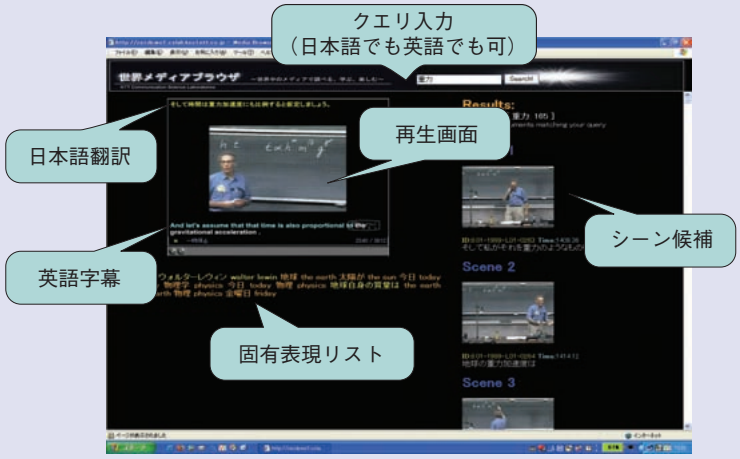


図1　世界メディアブラウザの画面

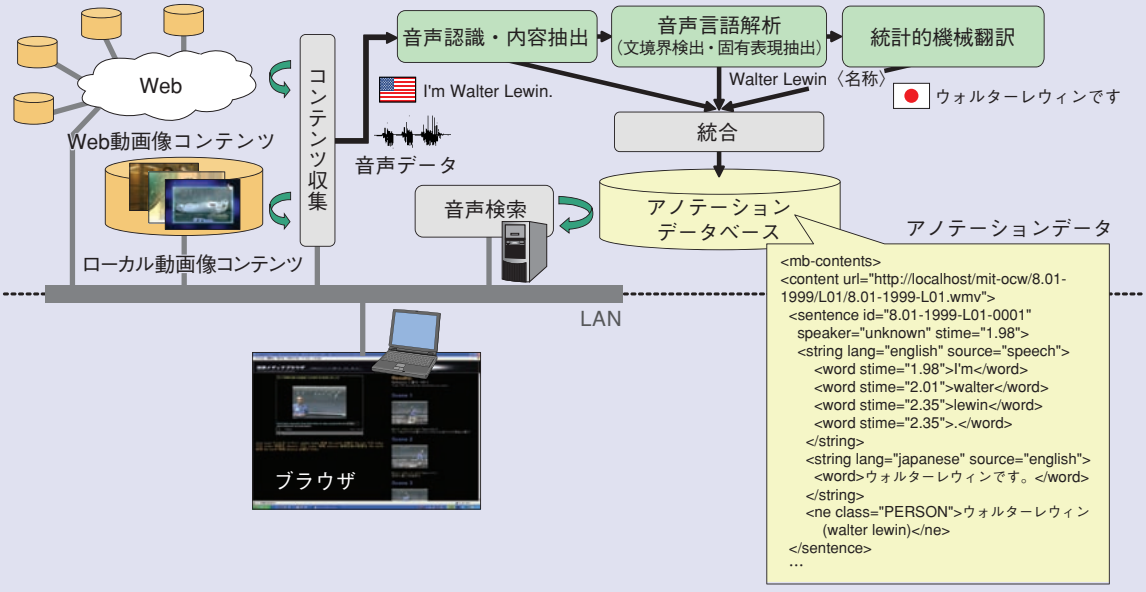


図2　世界メディアブラウザのコンテンツ解析

現在は，コンテンツとして実際にWeb上に公開されている米国マサチューセッツ工科大学（MIT: Massachusetts Institute of Technology）の講義収録ビデオ[(2)]を対象としてプロトタイプシステムを動作させています．

近年，大学の講義資料や講義収録ビデオをWeb上に公開するオープンコースウェア（OCW: Open Course Ware）の取り組みが世界的に広まっています．MITはその中心的役割を担い，多くの講義収録ビデオを公開しています．今後，このようなビデオコンテンツは世界中で公開されていくことが期待され，その先駆けであるMITのビデオは世界メディアブラウザの研究開発に適した研究試料といえます．

世界メディアブラウザを支える要素技術

世界メディアブラウザを実現するには，高い精度の音声認識，言語解析，機械翻訳の技術が不可欠です．なぜなら，音声認識の誤りが言語解析の誤りを誘発し，その誤りが翻訳の誤りを誘発するという具合に，小さな誤りが連鎖して拡大するためです．また，大量のデータを短時間で解析する技術や，任意のキーワードが含まれるシーンを素早く検索する技術も必要です．世界メディアブラウザには，私たちがこれまで研究してきた次のような成果が使われています．

■高速・高精度な超大語彙音声認識

音声認識は，音声信号をその内容を表す単語列に変換する技術です．入力された音声信号に対して，あらかじめ登録された単語の組合せの中から，音響的にも言語的にももっとも適合する単語列を見つけることができます．幅広い内容の音声を認識するには登録単語数を大幅に増やす必要がありますが，これは計算量の増加を招きます．私たちは，重み付き有限状態トランスデューサと呼ばれる計算モデルに基づいて，従来技術の数十倍の単語数（1 000万語）を扱える高速・高精度な音声認識アルゴリズムを考案しました．1 000万語という語彙の大きさは世界的にも例がなく，市販の音声認識エンジンは高々10万語程度です．また，一般的な国語辞典の収録語数は５～８万語，広辞苑でも23万語です．音声認識システムは，登録されていない単語を話されると，登録されている別の単語に誤って認識してしまいます．膨大な数の単語を登録しても高速かつ正確に認識できる本手法は，幅広い内容の音声を認識するうえで大きな強みになります．

また，音声認識に必要な音響モデルや言語モデルを構築するために，モデルの識別能力を直接向上させる識別的学習や，頑健なモデルパラメータ推定を可能とするベイズ学習を利用する方法も提案しており，高い認識精度を実現しています．

■高度な音声言語解析

音声言語解析部では，文境界検出と固有表現抽出を適用しています．文境界検出は，音声認識の結果を翻訳するうえで不可欠な技術です．音声認識結果には人が記述したテキストにみられる句読点は存在しません．そのため，翻訳の基本単位である文の区切り（文境界）を自動的に決定する必要があります．しかし話し言葉は，そもそも文という単位があいまいなため，適切な文境界を見つけることは簡単ではありません．私たちは，文末表現の特徴や文の構造（語と語の修飾関係）を利用した文境界検出手法を提案し，従来手法に比べて良好な検出精度を得ています．

一方，固有表現抽出では音声認識の誤りによる悪影響を最小限に抑える効果的な抽出方法を提案しています．実際に音声認識の誤りを完全に排除することはできません．そこで，音声認識時にどの程度の自信を持ってその結果を選択したかを表す確信度を，固有表現らしさを表すモデルの特徴に組み込みました．これにより，認識誤りに影響されにくい高精度な固有表現抽出を実現しました．

■階層的な句に基づく統計的機械翻訳

機械翻訳部には，統計的翻訳手法[(3)]を採用しています．統計翻訳手法は，専門家の言語知識に頼らなくても学習データさえあればあらゆる言語間の機械翻訳を可能にする技術です．学習データとしては，同じ内容を表す２カ国語の文の対を大量に集めた対訳コーパスを主に使います．統計翻訳は世界メディアブラウザのように多様な言語を扱うアプリケーションにおいては，必須の技術だと考えています．

私たちの統計翻訳手法は階層的な句に基づく手法を採用しています．この手法は，対訳コーパスから統計量でスコア付けされた文法を自動獲得し，これを翻訳モデルとします．**図３**は，獲得した文法によって得られる日本語と英語の階層的な句の対応関係を表しています．入力された原言語の文に対して，もっとも適切な句の対応関係を与える目的言語側の文を求め，翻訳結果とします．

この方法は，翻訳モデルの中に語の

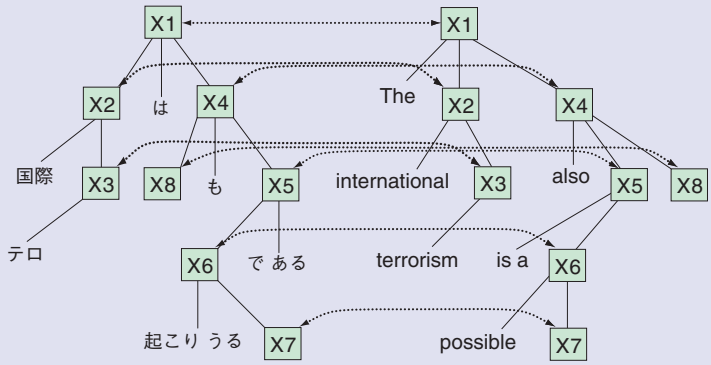


図 3　階層的な句に基づく統計的機械翻訳

並び替えモデルを統合した手法となっており，日本語と英語のように語順の大きく異なる言語間の翻訳で威力を発揮します．そして，目的言語側の言語制約を効果的に適用することで，高速な翻訳を実現するアルゴリズムも提案しています．

■オープン語彙音声検索

大量の音声データから任意のキーワードを含む区間を瞬時に見つけることは簡単ではありません．話された内容を100％正しく認識できれば，テキスト文書の場合と同様に，索引（キーワードとその出現区間を対にした情報）を構築することで高速な検索が可能です．しかし，実際には音声認識誤りを避けることはできず，誤った区間は正しく検索されません．さらに，検索キーワード（クエリ）が音声認識システムに登録されていない未登録単語であれば，そのキーワードを含む区間は必ず認識誤りとなるため，全く検索できない状況に陥ります．私たちは音声認識結果だけを保存するのではなく，音声認識時に正解に近いと判定された複数の候補を求め，それらをコンパクトに表現したコンフュージョンネットワーク（CN: Confusion Network）と呼ばれる構造に変換し保存します．そして，このCNから高速に検索するための索引を構築する方法を開発しました．CNは音声認識結果よりも正解を含む可能性が高いため，より正しく音声区間が検索されます．さらにクエリが未登録語であってもその発音情報を利用して検索する拡張を行いました．この拡張では，クエリが音声認識システムに登録された単語であれば索引中の単語，未登録の単語であれば索引中の発音と比較するため，登録単語，未登録単語，およびそれらが混在するクエリに対して高い検索精度が得られます．

今後の展開

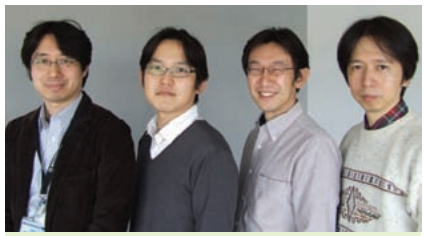
世界メディアブラウザの研究について紹介しました．現在，MITの講義収録ビデオを対象にした評価では，音声認識の精度は字幕としておおよそ内容が理解できるレベルを達成しています．音声認識を含む最終的な翻訳結果は，全く英語の分からない人が内容を理解できるほどの精度には至っていませんが，英語を部分的に聞き取れる人が理解の手助けとして利用できるレベルを達成しています．

音声認識から翻訳までの結果を詳しく分析したところ，音声認識や文境界検出のわずかな誤りが翻訳精度を大きく劣化させる例がみられました．また，話し言葉特有の流暢ではない発声や，非常に専門性の高い内容に対して，音声認識および翻訳の精度が低下することも数多く観察されました．

私たちは，これら課題の解決を目指して，現在も研究を進めています．そして，世界メディアブラウザが将来の通信放送連携におけるコンテンツ流通の核となる技術になるよう検討を重ねていく予定です．

■参考文献

⑴ 堀・須藤・大庭・渡部・小川・渡辺・マクダーモット・塚田・中村：“「世界メディアブラウザ」－音声認識と統計翻訳に基づく多言語動画コンテンツ検索／閲覧システム－，”日本音響学会講演論文集，1-1-17, 2008. 9.
⑵ http://ocw.mit.edu/
⑶ 塚田・渡辺・鈴木・永田・磯崎：“統計的機械翻訳，”NTT技術ジャーナル，Vol.19，No.6，pp.23-25, 2007.



（左から）堀　貴明/ 須藤　克仁/
塚田　元/ 中村　篤

音声・言語処理技術のブレークスルーと応用範囲拡大を目指し，これからも研究を進めていきます．

◆問い合わせ先
NTTコミュニケーション科学基礎研究所
メディア情報研究部
信号処理研究グループ
TEL　0774-93-5336
FAX　0774-93-1945
E-mail　hori@cslab.kecl.ntt.co.jp