Web時代の音声言語処理*

○中村　哲[1,2]，鳥澤健太郎[2]，隅田英一郎[2]，河井 恒[2]，柏岡秀紀[2]
[1]奈良先端科学技術大学院大学　情報科学研究科
[2]情報通信研究機構　ユニバーサルコミュニケーション研究所

1　はじめに

1991 年に欧州原子核研究機構 (CERN) のティム・バーナーズ＝リーが WWW: World Wide Web(以後 Web)に関する情報，ブラウザ等を公開して以来，Web には多様で膨大な情報が集るようになった．現在，Web には，全世界の全言語の情報（画像を含む）を加えるとゼタバイト（10 の 21 乗）オーダのデータがあると言われている．この中には，ビデオ映像，マルチメディアデータ，多言語の広告のページ，情報案内のページ，情報提供のページ，E-Commerce のページ，さらには，一般の利用者のブログ, 最近では Twitter などの情報が含まれている．Web データは多様な利用者によって日々生成, 更新されており, 日々変化していく実社会を射影した世界を構成している．この情報の大洪水の中で必要な情報を取り出すための中核的技術が言語関連技術である．

米国の情報関連研究施策を取りまとめている機関である NITRD[1] でも ”BigData”, “Human Computer Interaction and Information Management” を情報通信に関する 10 の重点課題の中に位置づけている．本稿では，音声言語処理の対象としての Web 情報, 音声言語処理を高度化する Web 情報, の 2 つの観点から現在の音声言語技術について考察する．

2　Web，ネットワークと音声言語処理

2.1　音声言語処理の対象としての Web 情報

Web には，日々変化していく実社会を射影した情報が存在する. ホームページ, ブログ, ニュース，Wikipedia や，最近では Twitter や Facebook などの位置，時間情報を含んだソーシャルメディアなどもあり多様化が進む．これらの情報の関連づけ，検索，提示などを行うためには，実社会の言語情報そのものを取り扱う大規模な処理系が必要となる．また，昨今ではビデオ動画のようなマルチメディアコンテンツが多く蓄積されており，音声処理も重要になっている．実際，インターネット上のデータ通信量では今やビデオ動画が圧倒的な量を占有していると言われている．

ビジネスの観点ではネット通販に代表される E-commerce が広く利用されている．レコメンデーション技術が有用であるだけでなく，これらのサイトは様々な商品に対する商品説明を伴っており，情報発信者が言語的アノテーションを自ら付与する点が興味深い．

このようなマルチメディアの Web コンテンツにアノテーションを付与する，相互に関連づける，検索し適切に提示するためのもっとも自然なツールが言語関連技術といえる．

もう一つの音声言語処理利用の観点は，これらの Web コンテンツの情報処理と利用者のインタラクションの高度化である．昨今，音声認識の語彙数，性能が向上したことにより，スマートフォンで音声翻訳（NICT の VoiceTra[2]），Web 音声検索（Google 音声検索など），音声対話（Apple の Siri, NICT の AssisTra[3]，質問応答システム一休[4]）などのサービスが登場し注目を集めている．

2.2　音声言語処理を高度化する Web 情報

膨大な Web コンテンは音声言語処理の高度化にも利用可能である．Web 上にある膨大な音声データ，テキストデータをクローリングして利用することで，音声認識の音響モデル, 言語モデルの高度化が可能となる. また，処理系をスマートフォンのような端末とサーバを接続した形態にすることで，多くの使用者からのデータを集約し集合知として利用することで,「使えば使うほど賢くなる」システムを構築することが可能となる．

* Speech and Natural Language Processing in the Web Information Era, by Satoshi Nakamura (NAIST/NICT),Kentaro Torisawa, Eiichiro Sumita, Hisashi Kawai, Hideki Kashioka(NICT)

3　研究動向

3.1　音声言語処理の研究動向

最近の米国の DARPA プロジェクトとしては GALE (Global Autonomous Language Exploitation, 2006-2011)プログラム[5]が有名である．このプロジェクトではアラビア語と中国語のニュース音声を自動認識し，英語への翻訳，情報抽出を行うもので（オフライン処理可能), これまでアナリストが人手で行っていた分析を自動化することを目標にしている．2012 年からは，さらに一般の中国語とアラビア語の各種方言の話し言葉を対象にリアルタイムで音声翻訳，情報抽出，検索処理するための BOLT (Boundless Operational Language Translation）プログラム[6]を開始する．このプロジェクトでは英語からの情報検索もターゲットに含んでいる．

3.2　情報アクセス技術

本稿では情報検索，情報抽出，要約，質問応答の４つの技術をまとめて情報アクセス技術と呼ぶ．他に評判抽出や文書分類などの技術もあるが，それらは情報抽出や情報検索の部分的な技術, 類似技術と捉えることができる．このような情報アクセス技術は，米国では 1980 年代から国防省や NIST の大規模プロジェクトとして遂行されてきた(Tipster，TREC, TIDES, GALE, AQUAINT, TAC, KDD など). 固有表現や質問応答といったタスクや，情報抽出の関係抽出，イベント抽出といった概念はこの流れの中で生まれてきたものであり，情報アクセスだけに限らず，翻訳や自然言語の基礎技術に対しても様々な影響を与えてきた．米国での情報アクセス技術に関する技術動向としては，固有表現などの限られた種類の分類問題に帰着される技術を機械学習手法で解くような研究はピークを過ぎ，対象が幅広いために，教師なしで知識を学習したり，クラウドを使って知識を作成し，Knowledge Bottleneck といわれる，知識作成の困難さを解決しようという方向の研究が盛んになっている．

4　NICT における取り組み

4.1　音声翻訳

4.1.1　多言語音声翻訳システム

現在主流の音声認識システムは，統計的音声認識手法を基礎としている．音声中の個々の音素の振る舞いや，単語の並びなどを統計モデルで表現し，様々な仮説の中から最も高い確率が与えられる単語列を認識結果として出力する．従って，これらの統計モデルの精度が音声認識性能に大きく影響する．NICT では，㈱国際電気通信基礎技術研究所（ATR）において長年にわたって収集してきた大規模音声コーパスに加え，インドネシア語やベトナム語などの多言語音声コーパスの収集を行い，多言語音声認識システムの研究開発を行っている．現在対応している日本語，英語，中国語，インドネシア語，ベトナム語の言語の音声認識システムの音響モデルは，表 1 に示す音声コーパスにより学習している．言語モデルの学習には, 日本語, 英語は 100 万文, 中国語は 50 万文, インドネシア語, ベトナム語は 16 万文の旅行会話文を用いた．

更に, 2009 年に行った全国 5 地方での実証実験期間中に収集された日本語約 6 万文，英語約 1 万 7 千文，中国語約 1 万 5 千文について人手による書き起しを行い，音響，言語モデルの両方について学習を行った[7]．また，VoiceTra 等で収集された書き起しの無い数百万文規模の音声データに対して，音声認識結果中の単語や文の信頼度が高い音声区間を用いて音響，言語モデルを再推定する，教師無し学習を行うことにより，音声認識性能の改善を図っている．

さらに, ユーザが自分に似た声で翻訳結果出力したいというニーズに対応するため，Web 上に大量に存在する音声コンテンツをクロールし，音質の高い声質変換が可能な平均声 HMM を自動的に学習する手法の研究を進めた．テキストが未知の音声に対して大語彙音声認識を適用し，誤りを含む音素列を参照して HMM の教師無し学習を行う場合，音素正解率が 80%程度以上であれば，良好な音質の合成音が得られることが明らかになった[8].

多言語音声翻訳の翻訳部では，全言語対に共通のシステム[9] を用いている．標準的なフレーズベース統計翻訳を基本として，いくつかの実用レベルの機能追加（固有名詞対訳の登録機能，翻字等）がなされており，旅行用の多言語対訳コーパスをモデル学習に用いている. 図 1 は, 旅行会話のテストデータ 500 文を用い, 20 の外国語から日本語への翻訳の評価実験を行った結果である．図に示すよう

に，当機構の翻訳システム（濃い灰色）は，広く利用されている多言語ソフトウェア（淡い灰色で表示）よりも高い翻訳率が得られることがわかる．

コーパスベース翻訳については，インパクトのある定説の一つが「量が質を決める」である．様々な実験から経験的に，「対訳コーパス量を増やせば翻訳品質が改善する」ことが分かっているので，対訳コーパスを効率的に収集することが重要になる．NICT は，日本語と外国語の対訳コーパス構築に注力しており，現時点で，日本語文とその対訳の対を単位として数えて総数 2700 万という世界最大規模に達している[12]．また，対訳コーパスは順次公開をしている．対訳コーパスの構築には，Web クローリング（Web データの自動収集）のようにコンピュータ中心のアプローチのほかに，外部機関との提携など，人中心のアプローチがある．前者の技術では，文書単位で対応する日本語と外国語のデータから

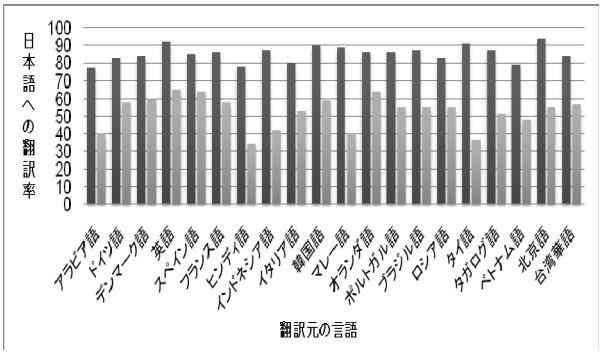

図 1　翻訳率の比較（広く利用されている翻訳ソフト（淡色）と NICT 翻訳（濃色）の比較．縦軸が日本語への翻訳率，横軸が翻訳元の言語）

文対応を自動的に抽出する技術が非常に有用であり，新聞，科学技術論文，特許など様々な分野の対訳コーパスの構築に活用してきた．また，後者については，翻訳のホスティング・サービス「みんなの翻訳」というユニークな試みを行っている[10].

表 1 音声コーパスサイズ

言語	話者数	文章数	発話時間	タスク
日本語	4,500	226,673	387.6	旅行会話
英語	930	236,737	257.6	旅行会話
中国語	540	213,352	251.0	旅行会話
インドネシア語	400	84,000	79.5	ニュース音声
ベトナム語	30	23,424	40.5	ラジオ音声

4.1.2　ネットワーク音声翻訳サービス[2]

音声翻訳技術の性能改善および周知を目的として，スマートフォン用の音声翻訳アプリケーション VoiceTra を開発し，2010 年 7 月 29 日より無料公開した．VoiceTra は，21 言語の双方向翻訳に対応しており，うち 5 言語については，音声による入出力が可能である．2011 年 12 月末時点での VoiceTra のダウンロード数およびアクセス数は 57 万および 650 万件となっている．VoiceTra は，ネットワーク型システムを採用しており，ユーザが発話した音声と翻訳結果はログとして音声翻訳サーバに蓄積される．表 2 に，音声ログ 100 件を無作為抽出し，聴取により内容を分類した結果を示す．約半数が旅行会話的な発話の翻訳に利用されていることが分かる．

表 2　VoiceTra の発話の分類

分類	比率(%)
無音	11
無効発話(非音声など)	11
明確な旅行会話	9
旅行会話と解釈可能	42
旅行会話以外の内容	27

4.1.3　国際標準化

音声翻訳の実現には，翻訳対象となる言語の知識および大規模な音声言語資源が必要である事から，一つの組織が全言語対，全ドメインの音声翻訳を実現する事は困難である．そこで，ネットワークを介して世界中に分散している音声認識，音声合成，翻訳モジュールを接続し全ての言語対を音声翻訳できるネットワーク型音声翻訳の実現を目指す．具体的には，ネットワーク型音声翻訳の実現に必要なモジュール間通信プロトコルとデータフォーマットの国際標準化が必要となる．

(1) アジアにおける国際標準化活動

アジアにおけるネットワーク型音声翻訳の先端研究を目的として，2006 年に ATR（日本），ETRI（韓国），NECTEC（タイ），BPPT（インドネシア），CASIA（中国），CDAC（インド）と共同でアジア音声翻訳先端研究コンソーシアム（A-STAR）を発足させ，2008 年には IOIT（ベトナム），I2R（シンガポール）が加盟して８カ国の研究機関と共同研究を行ってきた．2007 年にはアジア・太平洋電気通信標準化機

関（ASTAP）[11]にて標準化活動を開始し，2009 年 7 月，世界で初めてインターネットを介して異なるアジア言語を話す複数話者間で旅行対話を対象とした音声翻訳システムを用いて実時間音声対話に成功した．このネットワーク型音声翻訳技術をアジアに留まらず世界で用いられる標準化技術にすべく，標準化活動を ASTAP から ITU-T に移行した．

(2) ITU-T における国際標準化活動

2009 年 10 月 ITU-T の SG16, WP2, Q21 (Multimedia architecture)/ Q22 (Multimedia applications and services)において，ネットワーク型音声翻訳技術の標準化を始動した．NICT がエディタとなり，(1)ネットワーク型音声翻訳のサービス要求条件と機能，および，(2)アーキテクチャにおける要求条件の 2 件の勧告草案を作成し，2010 年 10 月 14 日に勧告 F.745 および勧告 H.625 として承認された[12].

4.2 高度言語情報融合フォーラム[13]

平成 21 年に，組織を越えて音声・言語の資源やツールを共有しつつ，言語の壁を感じさせない情報処理，コミュニケーションを実現するための技術の進歩発展・促進を図るために設立された．現在，民間企業（81 社），大学・研究機関及び国の関係者（158 者）が会員となっている．音声言語処理技術，情報検索や信憑性判定を含めた情報分析技術，これらの技術の前提となる今までにない規模の言語資源（辞書，コーパスなど）の研究開発，実証実験・標準化等を行い，その成果たるツールや言語資源を広く会員に提供すべく活動している．

5 まとめ

音声言語処理の対象としての Web 情報，音声言語処理を高度化する Web 情報という観点で、最近の研究動向を紹介した．みんなの翻訳や Amazon Mechanical Turk なども，インターネットにより実現した新たな音声・言語アプリケーションの方向と思われる。今後，インターネット，Web は音声言語処理だけでなく，AI 技術そのものにも新たな局面をもたらす可能性がある．一方で，音声認識の音響モデルや言語モデルの学習量と性能改善の関係を見ると，現在の学習データ量の範囲では依然性能が飽和しておらず，現在利用しているモデルの表現能力の不足も同時に痛感せざるを得ない．

謝辞

本稿は文献[14]を筆者らの執筆部分を元に再構成したものである。当該文献の共著者である乾健太郎氏，河原達也氏，黒橋禎夫氏，関根聡氏，木俵 豊氏，磯谷亮輔氏，堀 智織氏，松田繁樹氏の諸氏に心から感謝する．

参考文献

[1] http://www.nitrd.gov

[2] http://mastar.jp/translation/voicetra.html

[3] http://mastar.jp/assistra/index.html

[4] 鳥澤健太郎、情報爆発と音声アプリケーションの可能性 -言語処理研究者の考察-、情報処理学会研究会音声言語情報処理(SLP)、2010-SLP-84(17), pp.1-6, 2010

[5] http://www.darpa.mil/Our_Work/I2O/Programs/Global_Autonomous_Language_Exploitation_(GALE).aspx

[6] http://www.darpa.mil/NewsEvents/Releases/2011/2011/04/19_DARPA_initiates_overarching_language_translation_research_Publishes_Broad_Agency_Announcement_for_Broad_Operational_Language_Translation_program.aspx

[7] 河井恒，磯谷亮輔，安田圭志，隅田英一郎，内山将夫，松田繁樹，葦苅豊，中村哲, “H21 年度全国音声翻訳実証実験の概要,” 2010 秋季音響論集, no.3-9-6, pp.99-102, Sept. 2010.

[8] Jinfu Ni and Hisashi Kawai, “An unsupervised approach to creating web audio contents-based HMM voices,” Interspeech 2010, pp. 849-852, Chiba, Japan, Sept. 2010.

[9] Chooi-Ling Goh, Taro Watanabe, Michael Paul, Andrew Finch and Eiichiro Sumita, “The NICT translation system for IWSLT 2010,” IWSLT 2010, pp139-146, Paris, France, Dec. 2010.

[10] 内山将夫，阿辺川武，隅田英一郎，影浦峡, “みんなの翻訳,” 言語処理学会第 15 回年次大会論文集, pp.184-187, March 2009.

[11] http://www.apt.int/ASTAP-SNLP

[12] http://www.itu.int/ITU-T/studygroups/com16/index.asp

[13] http://www.alagin.jp/

[14] 中村 哲，磯谷亮輔，乾健太郎，柏岡秀紀，河井恒，河原達也，木俵豊，黒橋禎夫，隅田英一郎，関根聡，鳥澤健太郎，堀智織，松田繁樹，″Web 時代の音声・言語技術″, 電子情報通信学会誌 総合報告，Vol. 94, No. 6, pp502-517, 2011