1A3-OS-17b-1

Massive AI時代の音声言語技術

Speech and Natural Language Technologies in Massive AI Era

中村 哲[*1*2]
Satoshi Nakamura

[*1]奈良先端科学技術大学院大学
Nara Institute of Science and Technology

[*2]情報通信研究機構
National Institute of Information and Communications Technology

1. はじめに

近年、多言語音声自動翻訳の研究は大規模データと統計モデルにより大きな発展を遂げ、世界初の旅行会話音声翻訳ネットワークサービスとして運用が続いている。ネットワーク上の利用者が使うほどデータが集まり性能が高度化する、この技術はBig Dataに基づくAI技術と捉えることができる。本講演では、コミュニケーションのための音声言語処理、Web 上の多言語情報抽出のため音声言語処理の現状と今後について述べる。

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 でも”Big Data”, “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. 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. 音声認識の語彙数, 性能の向上、スマートフォンの登場で音声翻訳(VoiceTra[VoiceTra2009]), Web 音声検索(Google 音声検索など), 音声対話(Siri, しゃべってコンシェル、AssisTra[AssisTra2010], 質問応答システム一休[鳥澤 2010])などのサービスが登場している.

中村 哲、奈良先端科学技術大学院大学 情報科学研究科、奈良県生駒市高山町 8916-5 0743-72-5260, s-nakamura@is.naist.jp

2.2 音声言語処理を高度化する Web 情報

膨大な Web コンテンツ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.

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[Gale]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[Bolt]を開始した. 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.

4. 音声翻訳の研究

4.1 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, 多言語音声認識システムの研究開発を行っている. 現在の日本語, 英語, 中国語, インドネシア語, ベトナム語の音声認識システムでは、音響モデルは, 日英中国語 250 時間、インドネシア語、ベトナム語は 40～80 時間の音声データにより学習を行い、言語モデルは, 日本語, 英語 100 万文, 中国語は 50 万文, インドネシア語, ベトナム語は 16 万文の旅行会話文から学習している。しかしながら、このように事前に大規模なモデル学

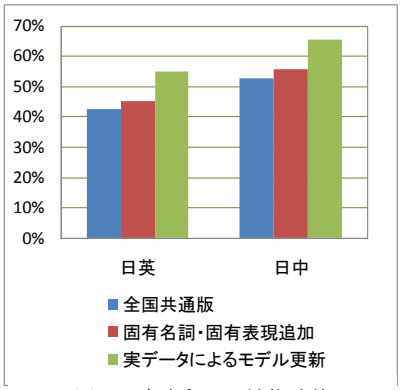

図１　音声翻訳の性能改善

習用のデータ収集を行うのは、収集のコストが膨大であること、実際の利用シーンとできるだけ同じデータが必要になることから、あまり効果的ではない。

2009 年に総務省の委託研究で全国 5 地方での音声翻訳実証実験が行われ、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実データ学習による評価が行われた[河井 2010]. 図1に音声翻訳の出力文の主観評価値（S:ネイティブ並み，A:申し分ない，B:まずまず，C:許容範囲，D:意味不明，の 5 段階で主観評価した際の S～C の比率）を示す。地域に応じた固有名詞、固有表現の追加と、実際の設置場所、応用システム形態での実データが性能改善を実現していることがわかる。

近年の多言語音声翻訳の翻訳部は，フレーズベースの統計翻訳が主流であり、言語対に共通のシステムを用いる．NICT[Chooi-Ling2010]のシステムでは、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2に，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を示す。NICT で開発した翻訳システム（濃い灰色）と併せて、Web で無料公開されている多言語ソフトウェア（淡い灰色で表示）との比較も表示している。

コーパスベース翻訳については，「量が質を決める」が定説になっている。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，現時点で，日本語文とその対訳の対を単位とし、総数 2700 万対という対訳文コーパスを構築している[12].

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 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がある[内山 2009].

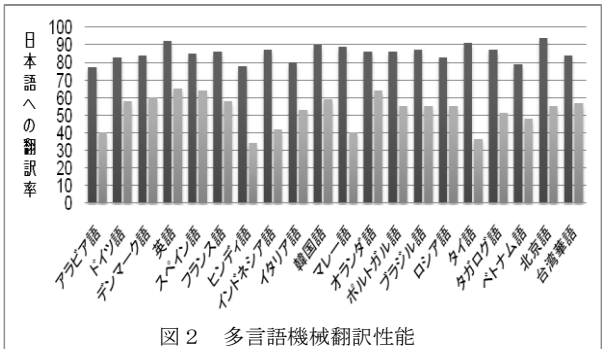

図２　多言語機械翻訳性能

4.2 ネットワーク音声翻訳サービス[9]

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 NICT において開発し，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約 60 万件および 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1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．

この VoiceTra で収集された書き起しの無い音声データに対しては，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，言語モデルのパラメータを再推定する教師無し学習を継続的に行っている。

5. まとめ

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．

表１　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

謝辞　本稿は文献[中村 2011]を筆者の執筆部分を中心に再構成したものである。共著者の皆様に心から感謝いたします。

参考文献

[VoiceTra2009]http://mastar.jp/translation/voicetra.html

[AssisTra2010] http://mastar.jp/assistra/index.html

[鳥澤 2010] 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

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

[Bolt]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

[河井 2010]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.99-102，2010.

[Chooi-Ling 2010] 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.

[内山 2009]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.

[中村 2011]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