マルチメディア通信と分散処理　62－18
（1993. 9. 21）

自然音声を認識･翻訳する技術

飯田　仁

ATR音声翻訳通信研究所
〒619-02　京都府精華町光台2-2
E-mail : iida@itl.atr.co.jp

音声認識、音声合成、言語翻訳などの個別技術の研究が成果を挙げつつあるが、自然音声を翻訳する観点から、現状の技術を眺めて、今後の問題を議論し、音声翻訳実現に向けた今後の研究動向を探る。

キーワード : 音声言語、音声認識、言語翻訳、翻訳電話

Atrs of Spontaneous Speech Recognition and Translation

Hitoshi IIDA

ATR Interpreting Telecommunications Research Labs.
2-2 Hikaridai, Seika-cho, Kyoto 619-02, Japan
E-mail : iida@itl.atr.co.jp

Research on elemental topics like speech recognition, speech synthesis and language translation has led to improvements in the accuracy and sophistication of each area of study. Viewed in translating spontaneous speech, we consider the state of the atrs in the areas from ponits of the future problems in developing a speech translation system.

Key words : spoken language, speech recognition, language translation, interpreting telephony.

1. はじめに

1990年代に差しかかるころから、画像、音声、言語のそれぞれの分野の認識·理解技術をより現実世界で通用する技術に仕上げていこうという機運が高まってきていると言えるだろう。つまり、現実世界における多様で、かつ動的な状況に耐え得る技術を目指して、これまでに達成し得た技術に課せられた制約を少しでも取り払おうと研究を進めているようである。

音声処理の研究について見れば、声の調子や老若男女に及ぶ不特定話者の音声認識は未だ大問題であるし、自然言語処理の研究について見れば、書き言葉と話し言葉とを区別してみたところで、いずれの文法も完全版には程遠い。このことは、記述すべき現実世界の状況が、非常に多様で、かつ状況に依存してその時点の状況の意味が決まるという動的可変な対象であることを示していると思われる。簡単な例を上げると、画像であれ、音声であれ、自然言語であれ、何らかのテンプレート·マッチングによる認識·理解の手法があるが、このテンプレートとはそのように認識·理解しようとした時に使える道具であり、認識対象がどのように変化していくか分からない時、つまりムーアの彫像であるかと思うとベンチであったり、またある時は橋であったりと変化する時、認識する側が何を見、聞き、解釈するかによりその対象が何者であるか決まってしまう。つまり、いつも家具しか現れないと考えていれば、橋を認識することができない。このように、認識·理解する側が外界世界の状況を可能な限り客観的にとらえる能力を備えることと、状況に応じて何を見て取れば良いかを判断できる能力を備えることが必要となってきていると言えるであろう。

さらに、今述べた各技術を統合していこうとする試みの一つとして、翻訳電話を目指す研究がある。現状は、音声認識、言語翻訳、音声合成の各技術を一方向に実行していくことで、翻訳結果を音声で出力するという方式を採っている。先に述べたように各技術の現状の限界がある一方で、翻訳の対象は声で発せられた話し言葉であるから、陳述の現れ方などの点で、文字で記述した「文字言語」とは異なることから、「音声言語」を扱う上での課題が多い。

本論では、まず2章で現状の翻訳電話システムにおける音声認識と言語翻訳の技術について述べる。3章では、音声処理と言語処理を融合するための手法について考察し、4章で音声言語の現象などを概説し、5章で音声翻訳の今後を展望する。

2. 翻訳電話システム

実用化システムという面では、それぞれの技術が日常環境下において十分に稼働するだけの基準に達していない。しかし、翻訳電話システムという総合的な技術目標を設定することにより、各要素技術の高度化が進められ、それらを効果的に融合していく技術が検討され始めた。

また、上記3要素技術の統合化には、現状の能力に適合したユーザ·インタフェースの機能も必要であり、翻訳結果モニター用の小型ディスプレイ装置の付加などの検討を要する。

2.1節で主な音声認識の手法を簡単に紹介し、2.2節で音声入力による言語翻訳の処理手法をいくつか紹介する。音声合成[1]については割愛する。

2.1 音声認識技術の現状

最近の音声認識研究の主な手法として、次の3種をあげることができる。

(SP1) 確率的音素モデルによる音声認識:

収録音声に基づいて音素の確率モデルを作成して、それらを単位にして入力音声の解析をする。このとき、単語を形成する音素列の現れ方を一種の文法と見なし、この文法に従って次に現れる音素の認識を高精度に行う[2]。

(SP2) ニューラルネットによる音声認識:

ニューラルネットを使って、音素識別および単語識別を行う[3]。

(SP3) スペクトログラムリーディング:

スペクトログラムを図形情報として扱い、そのリーディング知識を使ったエキスパート·システム[4]で、音素の分節化などに効果的である。

SP1の手法では、音素モデルとしてHMM(Hidden Markov Model)を使い、音素に関する予測制御として拡張LR構文解析手法で使われるLR-テーブルと呼ぶ解析の動作表を使う。この認識手法をHMM-LR法と呼ぶ[5]、[6]。LR-テーブルは、音声を文字表記するための写像を与えている。

その他、文脈情報を使った音声·言語処理の試みや、マーカ伝搬による音声·言語処理の試みなどがある。前者では、対話の展開などに関する文脈的状況を予め知識として記述しておき、単語認識の際の予測条件として利用する考え方である。MINDS[7]などのシステムで実験が行われている。後者では、ダイレクト·メモリ·アクセス解析法(direct memory access parser)という特殊な文解析の処理対象を、一方では語から音素へ、他方では文から対話対などの文章へ拡張していき、音声レベルから言語レベルまでを統一的に扱う手法が検討されつつある。多くの課題が残っているが、小規模ながらφDMDIALOG[8]、[9]などの実験が試みられている。

2.2. 言語翻訳

(LT1) 言語解析を中心とする翻訳手法:

翻訳電話実現のためには、書き言葉とは異なる言語現象を扱わねばならない。省略を含む断片的な発話をできるだけ正確に捉える方法や、発話意図に関する表現などを柔軟にかつ適切に捉える方

法を考慮すると、単に文脈自由文法で記述した文法を使うだけでは、処理できる言語現象に柔軟に対応できず、あるいは文法規則を詳細化すればその数が増大していく。このような状況では、制約に基づく表現が有効と考えられ、単一化操作に基づく語彙主導型の文法の枠組みが提示されている。この文法は、構文的な情報から、意味および言語運用論的な情報までを素性構造として記述でき、統一的に扱うことができる。この素性構造の単一化操作を行う解析機構の基本として、アクティブ·チャート解析法(10) を使う。この機構を使うことにより、音声認識結果の尤度をスコアとする最適解の早期抽出が可能となる。解析結果に従って翻訳する過程は、従来の変換方式と呼ばれる機械翻訳の手法に従う。このような手法を使った音声翻訳のシステムとしてASURA(11)がある。

(LT2) 対訳用例を利用する翻訳手法：

我々が外国語を学習する時、典型例文を暗記し、それを応用していく場合が多いと思われる。ビジネス·レターや対話での受け答えの慣用的な表現などを計算機で翻訳しようとする時、典型例文を応用することにより、翻訳を実現する手法である。そこでは、適切な例文を捜し出すために、文と文との間の類似性を計量する仕組みが導入される。このような手法を使った対話文の翻訳システムとしてTDMT(12)がある。

(LT3) 言語モデルによる確率的翻訳手法：

翻訳対象となる言語のモデルと、訳文の現れ方をモデル化する翻訳モデルとから翻訳言語対の間での文間の確率を求めることから始める。その確率は2言語間の大規模コーパスから求める。これが求まると、両言語間の複合器を想定し、与えられた目標言語の文に対して生起する確率最大となる原言語の文を翻訳結果とする。英仏間の大規模な実験が試みられている。

また、対話を翻訳する観点から次のような個個の言語固有の性質を考慮した翻訳を行う必要がある。

(TR1) 補助的述語の表現(例:(と)思う、(で)結構です)。

(TR2) 日本語の受給表現、英語における受益行為に関する動詞(例: give,offer,etc.)。

(TR3) 日英のアスペクト表現の違い、特に日本語では、論理的に冗長となるアスペクト句(例えば、無料です→無料となっています)、英語では、動詞固有の性質を考慮したアスペクトのタイプ(progressive、resultive など)。

3. 音声·言語のインタフェース

音声·言語の統合化の手法として確実性が高い手法は、今のところHMM-LR であろう。音声·言語処理のインタフェースとして、複数候補の文節を扱うために、文節ラティスを取り上げる。

文節ラティス：

HMM-LR 法のモジュールから尤度つきの文節ラティスを出力する。文としての可能な候補をすべて解析して、構文的、意味的に適切な候補を選別するには、計算量が膨大になる。そこで、文節間の係り受け関係を使った文節候補の事前絞りこみの過程が必要となる。

文節間係り受け関係：

文節間の係り受け関係を使って最適な文節候補を選択する一般的な手法がいくつか考えられている。それらは、係り受け関係の曖昧性を一般的に捉えたもので、曖昧性が比較的多く出現する場合に効率的な算法となっている。それに対し、現状の音声·言語の処理技術の下では、対象領域を限定することが不可欠である。そのような限定領域では、係り受け関係の曖昧性が比較的低いため、それらの算法を使うと、かえって無駄になる探索を増やしてしまう。

4. 音声言語の現象

翻訳電話の言語処理の対象は、朗読されるテキストなどではなく、人間同士の情報交換のための対話である。したがって、対話を収録·分析すれば明らかなように、対話の各発話には書き言葉とは違う話し言葉の様々な言語現象が現れてくる(13)。特に、

(D1) 断片的で書き言葉の文として不完全な発話、

(D2) 聞き手に対する待遇的配慮が現れた丁寧な表現や依頼などにおける間接的な表現、

(D3) 文脈を背景に備えた言語外の情報に依存して決まる意味を担った表現、

などが特徴的な現象であると言えよう。また、対話の開始や終了のための決まり文句を含む社会的に習慣化している表現などがある。さらに、音声入力であることから、発声のアクセントや発話中のイントネーションなども文脈に応じて使い分けられ、構文的·意味的に曖昧な表現となることを阻止している。ただし、現状ではこのような韻律に関する音声情報を十分認識することは難しく、言語処理に有効に使っていくことはできない。以下で、音声言語に特徴的な現象について触れておく。

音声言語に依存した意味の表現：

構文·意味的に曖昧な発話でも、イントネーションやポーズの位置によりその意味が確定する場合がある。特に、Hirschburg 等は英語のイントネーション·パタンや強調の位置が発話の意味を確定するために大きな役割を果たしていることを示した(14)。例えば、次に挙げるような項目について曖昧性除去の例を示した。

[発話行為の解釈]

“Can you . . .” の疑問文が依頼の表現である

(間接発話行為)か、単なる質問である(直接発話行為)かなど。

[修飾関係やスコープの解釈]

副詞'only'の修飾先の決定や、'<形容詞><名詞$_1$> and <名詞$_2$>'なる並列句における修飾関係の決定など。

[新旧情報の解釈]

強勢の置き方に従った新情報の決定など。

[談話構造の解釈]

'now'などの手掛かり語の役割決定など。

日本語においても、応答(「そうですねぇー」や「そうですねっ」の違い)の解釈、指示詞の確定(指示詞「あの」や間投詞「あのー」の違い)などがあげられる。

5. 音声翻訳の今後の展望

対話を適切に翻訳するために、対話内容の理解がある程度要求されることになり、文脈処理と総称される高度で、かつ計算量的に高価な処理に期待がもたれる。しかし、実際的なシステム上にこのような処理を追加するには、実用的な面でまだ多くの課題が残る。

現状では、事例主導型理解手法:Scriptや MOPSなどによる典型事例を中心にした理解過程の解釈、プラン認識型理解手法:対話参加者の発話に関するプランを認識して、対話内容の理解(15)および次発話の予測(16)などがある。

また、4章で見たように、言語理解における種々の曖昧性の解消には、字面だけでは無理があり、対話理解においてはイントネーション・パタンなどが果たす役割が大きい。それらを統合した言語処理を実現していくための一つのアプローチとして、理解状態を把握していく方法が考えられる。その理解状態を、情報が流れていく記憶状態の変化として捉える試みがある。そこでは、言語の解析レベルの情報から、対訳用例に関する情報、連想知識に関する情報、さらに韻律などの音声に関する情報などが、自律的に問題解決する能力を備えると共に、同時に記憶状態の状況を情報源にもなっていると考える(17)。

このような記憶状態を保持する翻訳モデルを考えていく時、情報の粒度と個々の情報が相互に協調しながら問題を解決する協調的な分散人工知能を念頭に置いておくことが必要と考えられる。

6. おわりに

通訳電話実現のための基礎的技術について述べ、今後の問題を考えた。概説した。いずれにしても、計算の理論などの基礎的研究成果と、システム実現技術とにはかなりの隔たりがあり、まだそのギャップが埋まっていない。膨大な計算量をこなすことも重要な要素であり、かつ個別の認識技術を統合するという新たな挑戦が始まっていると言える。

参考文献

(1)Klatt, D.: Review of text-to-speech conversion for English, J. Acoustical Soc. of America,Vol. 82, No.3 (1987).
(2)Lee, K-F. & H-W. Hon : Large-vocablary speaker-independent continuous speech recognition using HMM, Proc. of ICASP, pp.123-126 (1988).
(3)沢井、他:連続音声認識のための時間遅れ神経回路網を用いた音韻/音節スポッティング、電子情報通信学会、Vol.J-72-D-II、No. 8、pp.1151 - 1158(1989).
(4)畑崎、他:スペクトログラム・リーディング知識を用いた音韻認識エキスパートシステム、電子情報通信学会技術研究報告 SP87-19、(1988).
(5)北、他:HMM音韻認識と予測LRパーザを用いた文節認識、電子情報通信学会技術研究報告 SP88-88、(1988).
(6)北、他:SL-TRANSにおける文節音声認識、情処学会第39回大会 (1989).
(7)Young, S., et al : Layering Predictions: Flexible use of Dialog Expectation in Speech Recognition, Eleventh International Joint Conf. on Artificial Intelligence, pp.1543 - 1549 (1989).
(8)Kitano, H. : A Model of Simultaneous Interpretation: A Massively Parallel Model of Speech-to-Speech Dialog Translation, Proc. of the Annual Conf. of the International Asociation of Knowledge Engineers (1989).
(9)Kitano, H. : A Massively Parallel Model of Speech-to-Speech Dialog Translation: A Step toward Interpreting Telephony,Proc. of EuroSpeech89 (1989).
(10)松本:統語解析の手法、田中・辻井編:自然言語理解、第3章、オーム社(1988).
(11)Morimoto, T., et al. : ATR's Speech Translation System : ASURA,Proc. of EuroSpeech93(1993).
(12)Furuse, O., & Iida, H. : Cooperation between Transfer and Analysis in Example-based Framework,Proc. of COLING92(1992).
(13)飯田:自然言語対話の言語運用特性と対話処理の研究課題、人工知能学会誌,Vol.3, No. 4, pp.49 - 56 (1988).
(14)Hirshburg, J. : Proc. of The 3rd Workshop on Theoretical Issues on Natural Language Processing (TINLAP-III), pp.86 - 92 (1987).
(15)飯田, 有田:4階層プラン認識モデルを使った対話の理解、情報処理学会論文誌,Vol.31, No. 6, pp.810 - 821 (1990).
(16)Iida H., et al, "Predicting the Next Utterance Linguistic Expressions Using Contextual Information", IEICE TRANS. INF. & SYST., VOL. E-76-D, No.1, Jan. (1993).
(17)Iida H., "Prospects for Advanced Spoken Dialogue Processing", IEICE TRANS. INF. & SYST., VOL. E-76-D, No.1, Jan. (1993).