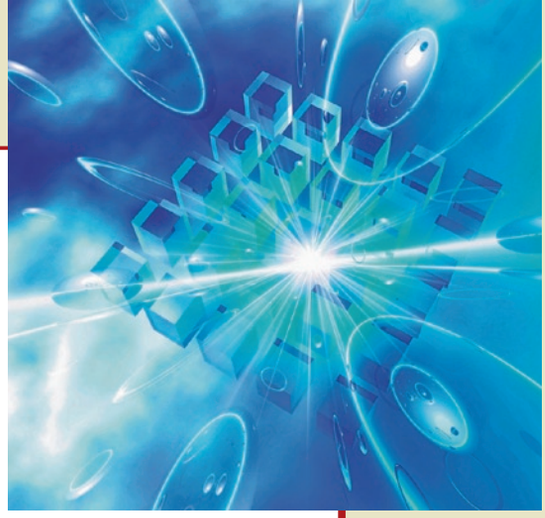

1 音声自動翻訳技術の進展

長尾　真（国立国会図書館）

音声翻訳に関する技術上の諸問題を論じた．まず音声翻訳の歴史について簡単に触れた後，音声の認識と合成の技術を紹介した．音声の解析には今日広く隠れマルコフモデル（HMM）が使われていること，音声の合成にはコーパスベースの方法が使われていることを述べた．次に言語翻訳の各種技術について紹介した．構文翻訳方式，意味翻訳方式，用例翻訳方式の特徴を述べ，統計的翻訳方式についても触れた．その後，会話文の持つ特徴や対話における問題点を指摘した．

自動翻訳電話の研究

自動翻訳電話を言い出したのは日本電気(株)で，ジュネーブで開催された 1983 年のテレコム '83 で将来の夢の通信として非常に簡単な音声翻訳のデモを行い注目を浴びた．しかし企業の研究所にとっても，このシステムの研究開発は負担の大きなものであり，何年先に商品化できるかはまったく予測できないという時代であった．自動翻訳電話の研究は，日常会話音声の認識，不完全な文の翻訳，その結果を発話文の形に音声合成して出すという 3 つの部分からなるが，それらのいずれの部分についても多くの困難があり，深い研究を必要とする．またこの種の研究は大きな辞書や巨大なデータを用いた大規模システムとなる．そういったことから，この研究開発を，大きな研究組織を持てない日本の大学で行うことは難しい．したがって，こういった研究開発こそ国の設立する電気通信基礎技術研究所（ATR）でやるべき研究であるとして，1986 年の設立から今日まで国からの研究費と一部民間企業の資金の投入によって研究が行われてきているのである．

テレコム '83 以降，このような動きに関連して，海外ではカーネギーメロン大学，カールスルーエ大学などが大きなチームを組んで研究を開始した．企業としては当時はブリティシュ・テレコムや ATT ベルラボなどが 1990 年代の前半に研究成果を少し発表するようになった．その後もいろいろな所で研究が行われ，今日では，いくつかの企業で音声認識や自動翻訳電話の装置が作られるようになってきている．しかし，それらも場面や話題，使用単語数や発話文体などをある範囲に限定しており，そういった条件下でなんとか使えそうだという技術レベルである．

音声の認識と生成

人間の音声の性質を調べる研究は 1950 年代から行われていた．音声の周波数分析から，フォルマントという母韻に特徴的な性質が存在することが分かり，これによって母韻の認識を行うことが広く行われた．また音声波形のパワースペクトルの対数変換を逆フーリエ変換したケプストラムによってピッチ抽出等を行う方法も開発され，その他種々の方法を駆使して音声波形の分析を行い，音節の認識を行うことができるようになった．

しかし音節の持つ性質は，発話者はもちろんのこと，発話における隣接音節によっても大きく影響されるといったことから，その取り扱いは確率的立場から行う必要があることが広く認識され，音声認識のほとんどすべての段階で確率モデルという考え方が導入されている．

入力音声を X，その認識結果を W で表そう．W が音節の場合には，X はそれを構成する音素（多くの場合，10 ミリ秒ごとの音波の性質（多次元の特徴ベクトル））の列であり，W が単語の場合は X は音節の列である．さらに W が 1 つの文である場合には X は単語の列と考えてよい．

そこで X が与えられたときに，P(W|X) が最大となる W を X の認識結果とするのが最尤推定法による決定であり，これは

$$P(W|X)=P(W)\cdot P(X|W)/P(X)$$

であるので，結局 P(W)・P(X|W) を最大にする W を探し出すことが課題となる．ここで P(W) は，認識の対象となる話題領域において W が現れる確率であり，

P(X|W) は W の発話によって X が生起する確率である．この考え方は 1980 年代から現在に至るまで音声認識の基礎となっている．

P(W) が信頼できる値となるためには，W の数が比較的少なくて，かつ P(W) を計算する発話データが大きくなければならない．日本語の音節は 100 個ほどであるから，それほど大きくないデータでも P(W) は計算できるだろう．しかし P(X|W) の計算にはさらに大きなデータ量が必要となる．発話データの中で特定の W の出現頻度はあまり大きくない．その W の特定の現れ方 X は，W の出現頻度のさらに数十分の 1，あるいはそれ以下ということになるからである．W が単語である場合には，何万語という単語を相手にしなくてすむように，対話の場面を限り，そこで使われる語をできるだけ少なくする努力がなされている．

まず音節の認識について考えよう．1 つの言語の音節の数は 100 個ほどであるから，P(X|W) を計算できなくはないが，音声はあまりにも個別的，変動的であるので決してこの確率が安定的に得られない．そこでまず考えられたのが，W の発話の音素列 X が W の標準的な音素列との間で最もよくマッチすることを探索するダイナミックプログラミングの方法であった．しかしこれも不特定話者，大語彙などになると適当な手法とは言えず，確率統計モデルである隠れマルコフモデル (HMM) が広く使われるようになってきた．これの詳細は他書にゆずるが [1]，音声の時間とスペクトル空間の両方での変動に対処することができるモデルで，しかも確率的パラメータを大量の音声データで学習していくことができるという利点を持っている．このようなモデルが不特定話者や連続音声認識に適用できるようになったのは，数百・数千時間にも及ぶ大規模な音声コーパスの整備に負うところが大である [2]．

この HMM モデルは音節の認識だけでなく，音節のつながった単語の認識にも使われるが，これは各音節の隠れマルコフモデルを連結したもので実現している．ただ認識すべき単語が多くなると，これらすべてについて隠れマルコフモデルを作るのは大変なので，単語の集合を音節の木構造に作り，単語の第 1 音から順にこの木をたどっていく方法をとることによって全体としてのモデルのサイズを小さくする工夫もなされている．

単語から文への認識については単語の N-gram モデルや有限オートマトンモデルが用いられている．これらは短い発話文のときには使えるが，長い複雑な発話文を認識しようとする場合には，言語学的により適した句構造文法モデルを使うことになる．その中でも発話文の単語生起の順序性を考えると文脈自由型句構造文法のグライバッハ標準形を用いるのがよいだろう．

これからスーパーコンピュータが簡単に使えるようになり，また発話データも巨大なものが集められ利用できる時代になると，音節の認識だけでなく，単語や文まで最尤推定法をそのまま適用して認識を行う時代がくるかもしれない．

翻訳がなされた後は，単語列を音声に直す音声合成の段階に入る [3]．音声合成には単語を音節列に直し，それぞれの音節に対応する音声波形を音声波形辞書から取り出してつないで単語の発音とするという方式が最もプリミティブなものとして考えられたが，非常に質の悪い音声しか作れない．そこで各単語に対応する標準的な音声波形を辞書に記憶させておき，これをつないで文の発話とする方法が行われた．しかし，この方法も各単語の波形と波形との接続個所におけるスペクトルや基本周波数の不連続性，その他の問題があるうえに，文の発話におけるアクセントやイントネーションの付加をうまくしないと聞いていて不自然であり，理解が容易ではない．

そこで単語などを単位として，それぞれの単位ごとに多数の発話データを記憶した大規模音声データベースを用意し，文全体に対して予測的に与えられる基本周波数や継続時間等に従って，データベース中から適切な発話データを選んで接続することで，全体ができるだけ自然に聞こえるようにするコーパスベースの方法が考案され，今日ではかなり良質の音声合成が実現している．

言語翻訳のモデル

場面限定などの制約のもとで，音声認識によって発話が漢字かなまじり文に変換される過程を経て，ようやく翻訳の対象となる文が得られる．

機械翻訳は 1995 年前後までは，ほとんどが句構造文法によって文を解析し，得られた句構造の木を翻訳の相手言語の句構造木に変換規則を使って変換し，そのあと相手言語の文法によって文の生成を行うという方式（これを構文翻訳方式という）をとっていた．これはチョムスキー（Chomsky）の句構造文法に基礎をおいた方法である．この句構造文法は形式言語理論の 1 つの分野で，文法を科学的立場から形式化する方法として出されたもので，コンピュータで言語を扱う人たちにとっては魅力的な枠組みであった．しかし，この枠組みで 1 つの言語のあらゆる可能な文を解析したり生成したりする文法を書こうとすると非常な困難に出会うことは実際に文法を書いてみると分かってくる．

日本語の構文解析に適したものとして係り受け解析がある．これは日本語の伝統的な文法であるが，日本語パーサ KNP が示すように，良い結果を出している．

もう1つの言語の記述法は，述語を中心に文をとらえる方式である．格文法といわれるもので，たとえば他動詞であれば，その動詞の動作の主体となるもの（主格），目的・対象となるもの(目的格)等を定め，ある特定の動詞に対して主格や目的格になり得るものはどのような名詞であるかを意味素を用いて規定するという方法がとられる．こういった記述をすべての動詞のすべての用法(同じ動詞でも用法によって表現する意味が異なる)に対して行った辞書を作る．そして文が与えられると，動詞を中心としてどの語が主語，目的語等になるかを単語の位置と意味によって決めるという形で文の解析を行う．この場合，名詞句などの構造の決定は句構造文法や次に述べる係り受け解析によって行われる．格文法の動詞辞書には原言語の動詞の格構造が目的言語の動詞のどのような格構造に対応するかを記述しておいて，相手言語への変換を行う．1982年から4年間で我々が行った科学技術庁の機械翻訳システムの研究開発はこの意味翻訳方式によった．

構文翻訳方式や意味翻訳方式のいずれの場合も文法規則を整備するのが非常に困難であるほかに，原言語の文構造を相手言語の文構造に変換する変換規則群を過不足なく作るのは至難の技である．そこでこのような問題を克服する新しい方式として用例翻訳方式が考えられた[4]．短い文の場合には対訳文対を多数用意しておき（用例対訳辞書），翻訳すべき文がこの辞書のどの文に似ているかを調べ，類似の文があればその翻訳文に合わせて翻訳する．

長い文の場合には適当な長さの句に分割し，それらの句を用例対訳辞書に入れ，それらの句が組み合わせられている構造を文法規則でとらえる．長い文の対訳を用例対訳辞書に入れることはスペース的にも困難だし，長い文になればなるほど，種々の異なった翻訳表現が可能となること，また翻訳すべき入力文との類似性の検出の機会が極端に減るという不利な条件が出てくるからである．したがって多くの場合，たかだか数語からなる句とその翻訳とを用例対訳辞書に記憶し，類似の句の翻訳はこれを参照して行われる．そして翻訳された句の相手言語での文への組み立てはその言語の文法規則によって行う．ここで使われる文法規則は文を構成する基本的な規則であるので，比較的少ない数の安定した規則群であり，構文翻訳方式におけるように膨大な数の規則数とはならないですむ．したがって用例翻訳方式は構文翻訳方式との折衷方式といえばいえる．

この方式の利点は，翻訳すべき文(句)と類似の文(句)が用例対訳辞書中に見つからなかったときは，人手で正しい訳を与え辞書登録すれば，それ以後は類似の文(句)を翻訳できるところにある．このような非常に単純な方法で学習が行われ，システムの翻訳性能を徐々に向上していくことができる．また数単語の句単位に適切な翻訳を与えるために，全体的に見て翻訳の質が他の方法に比べて良いという利点もある．

構文翻訳方式や意味翻訳方式では，翻訳がうまくいかなかったときに，解析文法，変換文法，生成文法，あるいは意味素の付与のいずれに問題があったかが簡単に分からないし，分かった後も，それをどのように変えれば改善につながっていくかの判断が非常に難しい．したがってこれまでの多くの研究開発は構文翻訳方式で苦労した後に用例翻訳方式に移ってきている．

最近は統計翻訳方式が世界的に流行するようになってきた[5]．この方法は大量の対訳テキストを統計的に解析することによって最も尤度の高い翻訳対となる単語列を取り出すことを中心とする方式であるが，言語的知識をいっさい使わないので，しばしば不自然な対訳句を取り出すことになって，日英のように言語構造がまったく違う言語間の翻訳にはあまり適当な方法ではない．そこで用例翻訳における句に当たるものを人手や自動で決めて，これらの句の並びに関する統計的性質を調べて翻訳する方向に変わってきている．したがってこれは用例翻訳方式に統計的観点を導入して大量の用例(対訳テキスト)からより良い翻訳句を選択しようとする方式と見ることができるだろう．

こういった翻訳方式のこれまでの発展を見ると，これからは用例翻訳方式と構文翻訳方式の適切な組合せを統計的立場からうまく行うという3つの方式の融合という方向に進展していくのではないだろうか．ただ会話文のように，省略が多く，また倒置など語順が状況に応じてかなり自由に変わる文の場合には，格文法の考え方に基づく意味翻訳方式が有効であるといえるだろう．たとえば

“大森，この急行停まりますか”

“この急行，大森停まりますか”

を正しく解析するためにはこのような考え方が必要になってくる．

会話文の特徴

普通の文章の読み上げ(朗読)とくらべて会話文においては次のような特徴が認められる．

(i) 発話のあちこちに意味のない音(不要音，あー，えー，…など)が入ることが多い．

(ii) 発話の途中に比較的長い無音区間が存在するし，文の終わりが必ずしも明確ではない．

(iii) 言葉の省略や倒置などが生じる．

(iv) 発話は必ずしも完全な文をなさず，途中で終わったり，言いたいことが途中で変わったりすることもある．
(v) 日本語の場合に，特に接続助詞などで，いくつもの文をつないで発話することが多い．
(vi) 会話文に特有のくだけた言いまわしがある．

こういった会話文の特徴を発話の中に検出するためには，そこに現れる語句や文法的特徴とともに，発話の音声的特徴(韻律やストレスの置かれる場所，その他)の微妙なところまでをとらえて判断することが必要となる．

発話の文体が平叙文でも，抑揚や強調を置く部分などによって疑問文になったり，命令を意図した文になったり，発話者の気持を伝えようとする文であったりする．たとえば“わかった↗”といえば“理解したか”という問い，あるいは念押しであるのに対し，“わかった↘”といえば“理解した”という場合と，“了解した”という場合がある．あるいはまた“もうそれ以上は言わないでくれ”といった気持を表すときにも使われる．こういった場合にどのような翻訳文にするのが適切か，またどのような韻律をつけて出すべきかは，相手言語の持つ性質とともに，それぞれの地域での文化的，習慣的なことが関係するので難しい問題である．

機械翻訳の立場からよく検討しなければならないのは，上記の(iii)の問題であろう．日本語では通常主語が省略されるし，目的語もしばしば略される．これらは多くの場合，直近の文中に存在することが多いが，稀にかなり遠く遡ったところに現れる．

代名詞の照応の場合も同様である．

A：“穫りたての魚です．これいかがですか．”
B：“ええ，下さい．”

という会話では，“これ”を単純に this と訳すのでなく，これは魚を指し，フランス語では男性名詞だから le で受けるといった判断が必要となる．B の応答は種々の語が省略されているので，これらを推定して復元してから翻訳しなければならないが，実際の会話の場合には，単純に“下さい”の直訳の“Donnez”という一言でも十分に通じるということもある．

日本語の場合，返事が肯定か否定かが不明確な場合がよくあり，その判断は難しい．

“あすオペラにいこうよ．”
“いいよ．”

と言うとき，この返事の抑揚の微妙な違いによって，行こうという場合だったり，行きたくないという意思表示だったりする．

会話文では“注意の焦点”(focus of attentions)に注目することが大切となる[6]．日本語の場合は通常動詞に近い名詞がそれであるのに対して，英語では通常文末にくる．たとえば，

東京へ車で行きます．
I go to Tokyo by car.

となる．発話においてストレスが置かれる単語と，その場所との関係に注意が必要である．

会話文において，さらに注意しなければならないのは，テンスとアスペクトであろう．現在と過去，未来とを正確に把握して翻訳しないと会話がちぐはぐになってしまう．アスペクトについても同様であって，たとえば期待しているのか，単に未来の予想を言っているのか，といったことがはっきり区別できる必要がある．たとえば英語の must は“しなければならない”と“に違いない”の 2 つの意味があり，“you must know”の場合は前者であるといった判断は詳しい文法的知識が必要となる．

Yes，no の使い方を日本人がよく誤るというのは広く知られたことである．

“まだ終わりませんか．”，“はい，まだです．”
You have not finished yet ？　No, not yet.

対話の特徴

対話についての言語学，認知科学的研究はオースチン(Austin)やグライス(Grice)によって 1980 年代に盛んに行われた．対話は話者 A の発話に対して話者 B がそれに関連して発話をする．これが協調的に行われる場合には次の 4 つの条件が成り立つというのがグライスの協調の原則である．

(a) 量の原則：求められている情報を過不足なく与える
(b) 質の原則：嘘や根拠のないことを言わない
(c) 関連性の原則：関係のないことを言わない
(d) 様式の原則：不明確，曖昧なことを言わない

話者 B が話者 A の発話に対して，これらの原則に反する発言をした場合には，A は B が協調的でないと考えることになるとされる．しかしたとえば，“京都に 1 泊したいのですが．”という発話に対して，(a) の原則に従って，すぐに“この宿はどうですか．値段は○○．場所は…．…です．”といって紹介することは対話になりにくいわけで，“どんな宿をお探しですか．”と曖昧な質問に対しては曖昧に応答せざるを得ないし，少しずつ情報を与えていくことも大切である．

人間と機械との対話においては，機械の側は多くの知識と推論機能を持ち，グライスの協調の原則や間接発話行為の問題などを考えた応答のシステムを作る必要があるが，自動翻訳電話の場合は，対面する人と人との対話であるから，こういったことは対話者が心得ていて，ほとんど問題とはならない．

ただ，自動翻訳電話システムの立場からすると，このような対話の持つ特徴を発話文の音声認識や文の理解のための予測に使い，認識の精度を上げることが考えられる．たとえば交差点で，“駅はどちらの方向ですか.”という質問があったとき，“あっちです.”，“こちらの方向です.”といった返答が予想されたとすれば，返答の音声認識における選択肢のパープレキシティをかなり減らせるだろう．比較的簡単な対話場面については**表 -1** のような対応が期待できる．

人と人が対面で自動翻訳電話を通じて会話をする場合には，音声認識や翻訳がうまくいかず聞き手が理解できなくても，“もう一度ゆっくり言ってください.”といったことが言えるし，その場の場面知識は人間が持っているので，人と機械との対話の場合よりも困難性は少ないと思われる．周囲の雑音をキャンセルして発話ができるだけ明瞭にできる技術，ポータブルな装置の中に巨大なメモリと高い処理能力を持ったコンピュータを入れられる技術の開発が大切である．

質問	—	返答
依頼	—	承諾 / 拒否
申し出	—	受諾 / 拒否
誘い	—	受諾 / 拒否
感謝	—	承諾 / 拒絶
評価	—	同意 / 不同意
非難	—	否認 / 是認
挨拶	−	挨拶

表 -1　対話における対応（文献 6）より）

参考文献

1）北　研二，中村　哲，永田昌明：音声言語処理，森北出版 (1996).
2）山本誠一：コーパスベース音声翻訳技術，電子情報通信学会誌，Vol.83, No.8, pp.604-611 (2000).
3）広瀬啓吉：音声合成技術，情報処理，Vol.38, No.11, pp.984-991 (Nov. 1997).
4）Somers, H. L. : Example-based Machine Translation, Machine Translation, Vol.14, pp.113-158 (1999).
5）Ney, H. : One Decade of Statistical Machine Translation : 1996-2005, Proc. of MT Summit X, pp.i12-i17 (Sep. 2005).
6）石崎雅人，伝　康晴：談話と対話，言語と計算 3，東京大学出版会 (2001).

（平成 20 年 4 月 14 日受付）

長尾　　真（名誉会員）
mngo@ndl.go.jp

1936 年生．京都大学工学部電子工学科卒業．1973 年同大教授．1997 年同大総長．2004 年情報通信研究機構理事長．2007 年国立国会図書館長．自然言語処理，画像処理，電子図書館．