大規模言語データベース収集法の一提案

菅谷史昭　金城由美子　竹澤寿幸　菊井玄一郎　山本誠一（ATR音声言語コミュニケーション研究所）

6M−01

A proposal of a large language data collection scheme

Fumiaki SUGAYA, Yumiko KINJO, Toshiyuki TAKEZAWA, Genichiro KIKUI, Seiichi YAMAMOTO

(ATR Spoken Language Translation Research Laboratories)

1　はじめに

コーパスベースの音声翻訳システムの一例としてATRが研究開発したATR-MATRIXシステム[1]の特長を分析すると、平均単語エントロピーの低い文では高い性能を示すが、一方エントロピーの高い文では性能は低い[2]．また、利用者側からの視点として、代表的な言い回しの入力に対しては高い性能を発揮するが、少しずれた表現を使用するとその性能が低い．第二の問題を解決する直接的な方法は、代表的な言い回しだけではなく、少しずれた多様な表現も収集することである．しかしながら、話者の自然な発声だけでは、代表的な言い回しが高い頻度で収集される傾向にあるため、多様な表現を収集することが困難である．そのため、多様な表現収集には、その表現を目的にした収集が必要となる．本論文では、多様な表現の収集法を提案する．多様な表現が収集できれば、単語の連接確率がより適切に付与されることが期待できるので、統計的言語モデルで生じる第一の問題解決にも寄与する．更に、言語翻訳においても、対訳コーパスの増大とともに、翻訳知識を自動獲得する統計翻訳の研究[3]が盛んになってきた．統計的な手法であるので、十分に大きな対訳コーパスが必要であると言われている．本論では、これらの問題意識を持ちながら、音声翻訳システムのより一層の性能向上を図るための大規模な言語データベース収集法について提案する．

2　ATRで収集した対訳コーパスのサイズ

ATRで収集した言語コーパスとして、バイリンガル旅行会話とモノリンガル旅行会話を例に、収集の概要について説明する．バイリンガル旅行会話では外国人旅行者とホテルのフロント係が異なる言語，具体的には日本語と英語を話し，二人は相手の言語を理解せず，音声翻訳システムを介して話しているとした．そのため，バイリンガル会話収集には日本語話者，英語話者，日英通訳者，英日通訳者の計4人が参加した．状況設定の違いにより，話者の役割を旅行者またはフロントとした．双方向の同時通訳的な翻訳も可能な会話参加者の配置となっている．4名の会話参加者の発話はすべて録音され，書き起こされた．会話は現実世界の状況を反映することが望ましい．そのため，話者に会話の間はプロットにしたがってそれぞれの役柄を演ずるよう要請した．特にホテルのフロント係にはそれに類する経験のある人を配した．一方音声翻訳システム用の音声認識技術の研究開発には，より多様な音声データが必要になると考えられた．そこで，バイリンガルの制約だけを外した日本人同士のモノリンガル旅行会話を収集した．表1にバイリンガル旅行会話とモノリンガル旅行会話データベースの規模と特性を示す．いずれのコーパスも10Kオーダの発話を集めるのに、約1年間をかけて収集していて労力は大きい．

次に、パープレキシティと1文当たりの単語数から、文数の概略を計算する．テストセットのパープレキシティは20程度であり、1文当たりの単語長は11単語程度であるので、$20^{12}=4.096*10^{15}$と大きくなる．べきが11+1と1を加えているのは、パープレキシティの計算では、文末マークをカウントするためである．音声翻訳システムを介して対話実験を行うと6単語程度と短くなることが知られている[4]．全ての状況で有効な統計量であるかは不明であるが、この値を使うと、$20^{7}=1.28G$となる．上述のデータベースの文数との差は5桁となる．音声認識においては、音響モデルの助けがあるので、計算上の文数全てを集める必要はないし、言語翻訳においても翻訳知識には汎化性があるので、上限の文数の収集は必要ないと思われる．しかし、データ量を増やせば、言語モデルなどの性能が向上することから、データベースのサイズ拡大の要求は高いので、次節では効率的な収集法を新たに提案する．

表1　ATR言語コーパスの一例

会話形式	バイリンガル (J to E)	モノリンガル (J to J)
収集会話数	618	892
異なり話者数	71	499
発話総数	16,107	22,874
日本語形態素延べ数	301,961	491,159
パープレキシティ	18.4	21.4

3　セル形式言語データ収集法

多様な表現の収集を日本語表現収集について、表2の収集結果を使い説明する．日本語の多様な表現を集めるために、英文を作業者に見せる．この英文を種英文と呼ぶ．日本語を見せないのは、多様な表

表2 収集データの一例

How many hours will you be late?			
何時間	ー くらい ぐらい ほど 程度	遅れ 遅くなり	ますか そうですか
どれ	くらい ぐらい ほど	遅れ 遅くなり	ますか そうですか
どの	くらい ぐらい 程度	遅れ 遅くなり	ますか そうですか
何時間	ー くらい ぐらい ほど 程度	遅れる 遅くなる	予定ですか 見込みですか
どれ	くらい ぐらい ほど	遅れる 遅くなる	予定ですか 見込みですか
どの	くらい ぐらい 程度	遅れる 遅くなる	予定ですか 見込みですか

現の収集にあたり、日本語表現に引きずられること懸念しているからである．表2の例では、"how many hours will you be late?"が種の英文である．作業者は、この日本語翻訳を考える．翻訳結果は、表2に示す表形式で作業者が直接タイプ入力する．タイプにあたっては翻訳作業と入力作業を兼務するオペレータと、翻訳作業だけを行う3人の作業者の計4人で作業を進めた．データのまとめ方は、作業者にまかせている．行方向の複数のセルに文が対応する．左から右にセルを移動しながら、一つのセルに複数の表現がある場合は、セルに含まれている各ラインの表現を選択可能である．"ー"はラインが空白であることを示す．音響データは収集していない．例えば、表2の最初のセル列からは"何時間遅れますか？"、"何時間遅れそうですか？"などの日本語表現がとりだされる．最初のセル列で 1*5*2*2=20 通りの表現が得られる．対話を模擬した収集では、表現のバリエーションを集めるには、話者数や収集回数を増やして行う必要があるが、本提案手法によれば、表現をセル形式に整理しながら、集めているので、表現の抜けが少なく、また、セル形式は作業者の負担が小さな方法となっている．

4 収集実験

種英文に対する日本語の数を拡大率と呼び、拡大率を2のべき乗に丸めた頻度分布を図1に示す．126文の種英文は旅行会話文からランダムに取りだした．図1によると、平均拡大率は16であり、最大12,898

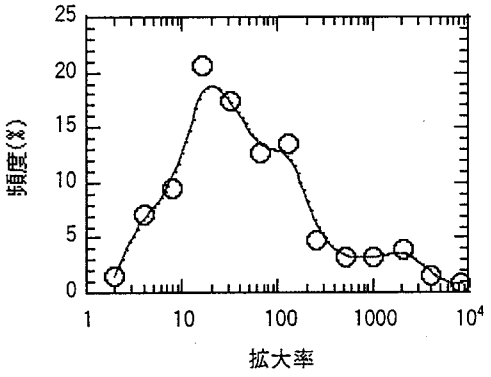

図1拡大率の分布

通りとなる．最小は種英文"Good night"に対する"おやすみなさい"あるいは"おやすみ"である．平均拡大率は436.9となった．

5 むすび

コーパスベースの音声翻訳システムの性能を改善するために、セル形式言語コーパス収集法を提案し、小規模な実験結果について述べた．拡大率が436.9となることから、同義な表現ながら、日本語文を2桁増やすことができることを確認した．種英文のサイズを更に大きくする計画である．言語モデルや言語翻訳に適用した場合の効果についても今後の課題である．

文献

[1] T. Takezawa, T. Morimoto, Y. Sagisaka, N. Campbell, H. Iida. F. Sugaya, A. Yokoo, S. Yamamoto, "A Japanese-to-English speech translation system: ATR-MATRIX", Proc. ICSLP 1998, pp. 2779-2782.

[2] 菅谷史昭，竹澤寿幸，横尾昭男，山本誠一，"音声翻訳システムと人間との比較による音声翻訳能力評価手法の提案と比較実験"，信学論(D-II)，vol. J84-D-II, no. 11, pp.2362-2370, Nov. 2001.

[3] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, "The mathematics of statistical machine translation: parameter estimation", Computer Linguistics, Vol. 19, No. 2, pp. 263-311, 1993.

[4] 菅谷史昭，竹澤寿幸，横尾昭男，山本誠一，"音声翻訳システム(ATR-MATRIX)の評価"，信学技報，SP2000-21，pp.39-45,June 2000.