E-016

種会話からの派生による話し言葉コーパス構築

Building Spoken-Language Corpus by Text Derivation from Germ Dialogs

麻野間直樹† **山田節夫**† **古瀬蔵**† **奥雅博**†
Naoki Asanoma Setsuo Yamada Osamu Furuse Masahiro Oku

１．はじめに

良質の話し言葉コーパスを低コストに構築し，日常会話をタスクとする音声対話システムの言語モデルを作成するために，文作成の基となる会話（「種会話」と定義）から新しい会話文を派生させる話し言葉コーパス構築の手法を提案する．

統計的な音声認識処理や機械翻訳処理に用いる良質な言語モデルを作成するためには，その利用場面によく適合したコーパス，すなわち利用場面で出現しうる文を大量に含んだコーパスを入手することが必要である．

しかし，話し言葉を対象とした言語モデルの作成を想定するとき，既存の話し言葉コーパスのうち十分なサイズのものは少ない．言語モデルを用いる音声対話システムの利用場面に適合するものという条件を加えると，既存の話し言葉コーパスから入手するのはさらに困難となる．利用場面に適合した既存の話し言葉コーパスが見つからない場合，従来は次のようなアプローチで新たに話し言葉コーパスを構築していた．

(a) 音声書き起こし：利用場面を想定して二人の話者が演じた会話を音声収録し，収録した音声をテキストに書き起こす[1][2].

(b) 作業者の内省：想定された利用場面で起こりうる細かな場面・条件をあらかじめ整理，細分化しておき，作業者がその一つ一つの場面を想像しながら，発話しうる文を机上で創作する[3].

(a)の方法は，(b)に比べてシステムが使われる状況に近いので，利用場面における言語表現をより多く含む良質なコーパスの構築が期待できる．その反面，収録時に二人の話者を同時に拘束する必要があるので，十分なサイズのコーパスを得るためには高いコストがかかる．

一方，(b)の方法は机上作業であるので，1 文あたりの作成コストは(a)よりも低く，同じコストならばより大きなコーパスを得ることができる．しかし，作業者が利用場面で起こりうる多様な表現を偏りなく継続的に作っていくのは，実際問題として難しい面がある．さらに本稿で対象としている日常会話のタスクは，あらかじめ利用場面を細分化し規定することが難しいので，場面の細分化を必要とするこの方法をそのまま使うことはできない．

このような問題点を解決するため，本稿では従来法(b)の低コストである利点を活かしながら，(b)を改良した種会話からのコーパス構築手法を提案する．

２．種会話からの派生によるコーパス構築手順

種会話から会話文を派生させることによって話し言葉コーパスを構築する提案手法の手順を以下に示す．

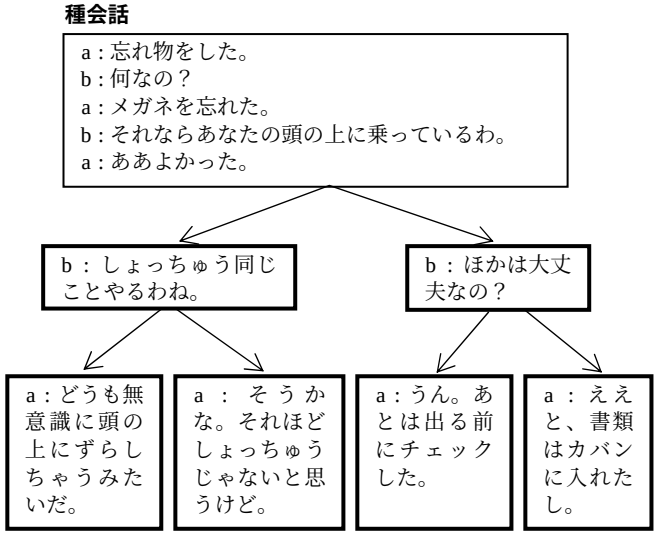

図１：種会話からのコーパス構築手順

手順１：最初に少量の種会話を用意する．種会話は，既存の会話集から引用したり，前述の従来のコーパス構築方法(a)を用いたりして用意することができる．種会話を選ぶ際の条件としては，言語モデルの利用場面で出現する文体・言い回しになるべく近いものがよい．

手順２：種会話を基にして作業者が机上で作文を行う．実際の作業としては，種会話の話の流れに合うように，種会話のあとに続く会話文を創作する．ここで，ある発話の次に続く発話を複数通り作っていくことで，会話の流れを分岐させ，作成文数をねずみ算式に増加させることができる．図１は，ある１つの種会話に対し，１ターン分つまり往復の２発話を，２分岐しながら創作する様子を表している．この場合１つの種会話に対し，図中太線で囲まれた６発話が得られることになる．

手順３：それぞれの種会話ごとに，ある定めた会話のターン数（すなわち分岐の深さ）を満たすまで，手順２を繰り返す．

この方法の利点は，少量の種会話さえ用意すれば，日常会話の中で起こりうる細かな会話の場面を考える必要がないことである．また作業者の観点からは，種会話の話の流れから会話場面を明確に想像することができ，会話の展開を想像しやすく，さらには創作作業を継続しやすい．したがって，日常会話のようなタスクに対しては従来法に比べて適した方法であると考えられる．

他には，種会話に倣って会話文を創作するので，話し言葉コーパスのトピックや文体などのコーパスの性質をコントロールしやすくなるといった特長もある．

† 日本電信電話株式会社　NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories

３．日常会話の話し言葉コーパスの構築

日常会話をタスクとする音声対話システムを構築するため，音声認識に用いる言語モデル用に，提案手法で日本語日常会話の話し言葉コーパスを構築した．コーパス構築の効率性を比較するために，従来手法として，音声書き起こし（1 節の(a)）と同様の手法でも，同じく日本語日常会話コーパスを構築した．

(i) 種会話

種会話は，市販の日常会話用例の本から５０会話を様々なトピックから抜き出して選択した．種会話の長さは，会話場面の話の流れがわかる程度の発話数として，２，３ターン程度にした．

提案手法に基づくコーパス構築作業を１０人の作業者で行った．１人の作業者は，１つの種会話について２ターン分（４発話）を，２分岐しながら創作することとした．これにより５０の種会話からは，50×(2+4+8+16)×10=計 15,000 発話が得られる．１つの発話に複数の文が含まれるので，全体の文数は発話数より多い 18,059 文となった．

(ii) チャット

従来手法によるコーパスは，初対面あるいは知人同士の２人がテキストチャットの画面上で雑談を行い，そのチャットの記録を日常会話の話し言葉コーパスとする方法で構築した．音声収録の条件に近づけるように，チャットを行う際は，音声で聞き，音声でしゃべっているつもりで会話を行い，文体も話し言葉口調になるよう，会話者に事前に指示した．この手法により 1,464 発話，合計 2,753 文を収集した．

なお提案手法（種会話），従来手法（チャット）どちらにおいても，次節の評価コーパスの内容を参照することなくコーパスを構築した．

４．コーパス構築の効率性評価

提案手法によるコーパス構築の効率性を，構築したコーパスから作成した言語モデルによって評価した．

良質なコーパス，すなわち対象タスクによく適合するコーパスを**より低コスト**で構築できる方法が，コーパス構築の効率が高いという考えに基づいて，3 節で構築した２つの日常会話コーパスの評価を行った．ここでのコストは，同じ人数の同じ能力の人が作業したのべ時間とみなす．

本稿ではコーパスの品質を測るために，それぞれのコーパスから作成した単語 trigram の統計的言語モデルにおいて，対象タスクに対する適合度を測る．

この対象タスクと言語モデルの適合度は，音声認識などの分野で伝統的に使われるカバー率や perplexity の指標を用いた．一般に，カバー率が大きければ大きいほど，perplexity が小さければ小さいほど高品質の言語モデルと考えることができる．

対象タスクは日常会話なので，評価コーパスとして，一般に手に入れることのできる LDC の CALLHOME Japanese Transcripts (LDC96T18) を用いた．実験は，CALLHOME のコーパス内のタグ情報などを除去して得られた 25,822 文をテスト対象とした．

効率性の比較を行いやすくするため，提案手法（種会話）のコーパスからランダムに 2,753 文を選び，従来手法（チャット）のコーパスの文数に揃えて評価した．

表１は，CALLHOME コーパスに対する，各言語モデルのカバー率，perplexity，2,753 文を提案手法によって作った場合のコストを１とした時の相対コスト，およびコーパスの文数を示している．カバー率はそれぞれの言語モデルと CALLHOME との間で計算した値，perplexity は，２つの評価コーパスの語彙を統合した語彙セットを用いて比較可能にした計算値である．なお，この統合した後の語彙数は 5,893 であった．

表１：CALLHOME に対する各モデルの適合度の比較

	カバー率	perplexity	相対コスト	文数
種会話	84.0%	513.9	1	2753
チャット	82.8%	667.0	4〜5	2753

表１より，提案手法による言語モデルは従来手法によるそれよりも良く（高カバー率，低 perplexity），なおかつコストが大幅に抑えられていることがわかる．したがって，提案手法によるコーパス構築は従来手法よりも効率性が高いと言える．音声書き起こしの作業を実際に行う場合，今回検証したチャットの作業に書き起こし作業が加わるので，従来手法のコストはより高くなり，効率性の差もさらに大きくなると考えられる．

なお perplexity が比較的大きな値になっている理由としては，もともと評価コーパス CALLHOME Japanese がくだけた表現を多く含んでいる一方で，今回実験した提案手法や従来手法のコーパスは整った丁寧な表現が多く，文体や言い回しなどの傾向や語彙そのものの分布に差があったためと考えられる．

５．まとめ

本稿では，低コストで良質な話し言葉コーパスを構築するために，種会話から会話文を派生させるコーパスの構築法を提案した．その効率性を対象タスクに対する言語モデルの適合度と作成コストで測ることとし，従来手法との比較実験を行った．その結果，提案手法におけるコーパス構築は効率性が高いことがわかった．

本稿で用いた種会話は作成するコーパスの質に強く影響する可能性が高いので，今後は種会話の設定の方法を詳細に検討する予定である．さらに実際に大規模な話し言葉コーパスを構築する際の提案手法の効果を測ることや，異なるコーパスに対する重み付け混合を用いた言語モデルで評価するなど，より詳細な評価を行う予定である．

参考文献

[1] 田中，他．自然言語処理－基礎と応用－，第 10 章，電子情報通信学会，1999.

[2] 竹澤，“音声翻訳研究用バイリンガル旅行会話データベースの構築”，言語資源の共有と再利用シンポジウム，1999.

[3] 甘粕，他．“サイバーアテンダント－自由発話入力に対応したマルチモーダル対話システム”， 日本音響学会 2004 年春季研究発表会講演論文集，3-Q-32，2004.