

6L-3 対話テキストの意味空間における分類

坂野俊哉 森元 逞
ATR自動翻訳電話研究所

1. はじめに

良質な言語処理のためには良質な言語モデルが不可欠である。言語モデルは言語現象を形式化したもので、より高品質なものとするためには、より多くの言語現象を把握する必要がある。そのためには、言語現象としての言語データを大量に収集することは勿論のことだが、際限ない言語データをいかに処理するかが問題となる。ATRにおいても自動翻訳電話の実現に向け、幾つかの言語モデルを作るために数多くの言語データを収集している。この言語データは、電話またはキーボードを用いた会話からなる対話テキストデータであり、対話データベースADD(ATR Dialogue Database)としてデータベース化されているが[1]、その効率的な分析が課題となっている。

2. 言語データの分類

大量の言語データをそのモデル化のために効率良く分析するためには、そのバリエーションで分類する方法がある。これまでに多くの種類の分類基準(バリエーション)が提案されてきた[4]。しかし、いづれの分類基準にしても、言語データが定量的に扱えなければならず、言語データ自身を数量化する必要がある。これに関しては、水谷[2]、西村[3]、Biber[4]などが実際に数量化理論に基づいて言語データを数量化し、言語のバリエーションの解析を行っている。今回、我々が用いているADDのバリエーションを解析すべく、ADD内の言語テキストの数量化・分類を試みた。この数量化には数量化理論第IV類および第III類によって行った。以下、順にその概要と結果を述べる。

3. 第IV類による言語データの数量化

数量化理論第IV類によるデータの数量化は簡単に以下のようなものである[2]。対象としているデータ同士が、どれくらい似通っているかを表す親近性を定義し、似ているデータならばなるべく近く、逆に似ていないデータならばなるべく離れて、1次元軸上にデータを配置するための手法であり、数学的には上記の親近性より求めた各データ間の親近度を計算し、それから導かれる対称行列の固有値・固有ベクトルを求めることに帰着できる。

3.1 テキスト間の距離

この手法では、親近性をいかに定義するかが重要となるが、我々はまず言語テキスト間の距離 d を次のように定義した。T1、T2を言語テキストとすると、

$$d(T_1, T_2) = \left[\sum_{w \in T_1 \cup T_2} \{P_1(w) - P_2(w)\}^2 \right]^{1/2}$$

とした。ここで P_1 、 P_2 はそれぞれ言語テキスト T_1 、 T_2 の中での単語生起確率を表す。この距離は単語の生起を基準として定義されており、0から1までの値となる。例えば語彙順に関係なく、テキスト T_1 、 T_2 の語彙構成が全く同じならばその距離は0となる。また、ADD内のテキストの長さがまちまちであることから、定義される距離はテキスト長に左右されにくいものが望まれるが、上記の距離 d から計算されるテキスト間距離とテキスト長との相関がほとんどみられない結果を得ている。この距離を基に、言語テキスト間の親近性 e を

$$e(T_1, T_2) = -d(T_1, T_2)$$

と定義した。

ファイル	会議	交通	参加	宿泊	案内	観光
1	○				○	
2	○		○			
3	○	○		○	○	
4		○		○	○	○
5		○		○	○	○
6	○		○			
7	○			○	○	
8	○		○			
9	○		○			
16	○		○		○	
17	○		○			○
18	○		○	○		○
19	○		○			
20	○		○			

表1 ファイルの内容 (10~15は省略)

3.2 実験

実験はADD内のキーボード会話テキスト15ファイル、電話会話テキスト5ファイルの計20ファイル用いた。ファイル1からファイル20までの内容を表1に示す。数量化した結果、固有値の大きい方から順に第1軸、第2軸、第3軸を取り出し、第1軸と第2軸からなる空間を図1に、第1軸と第3軸からなる空間を図2に示す。この結果から、各軸の表す意味はおおよそ次の通りであろう。

[第1軸] 観光案内の要素、事務手続きの要素の強弱を表す。

[第2軸] 電話会話の要素、キーボード会話の要素の強弱を表す。

[第3軸] 手続き・書類関連の要素、支払い関連の要素の強弱を表す。

この中で、第2軸による電話会話ファイルとキーボード会話ファイルがきれいに分離されているのが目につく。これは、キーボード会話の持つ推蔽され雑音が少ない書き言葉に近い特性と、電話会話の間投詞などの雑音が多い話し言葉に近い特性が反映されたものと考えられる。

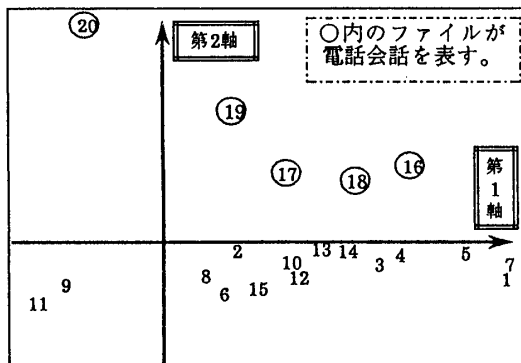


図1 第1軸と第2軸による空間

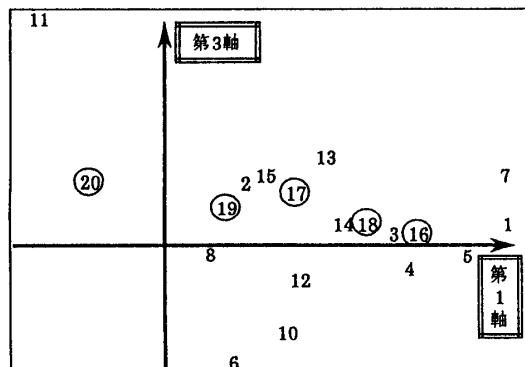


図1 第1軸と第3軸による空間

4. 第Ⅲ類による言語データの数量化

数量化理論第Ⅲ類によるデータの数量化は以下のように行う[5]。まず、幾つかの特性(特性項目)を決め、対象としているデータ(個体)がどの特性を満たしているかをマトリックスに表現する。第Ⅲ類は、満たしている特性のパターンが近いデータ同士をなるべく近くなるように、1次元軸上にデータを配置するための手法であり、数学的には、先ほどのマトリックスから導かれる対称行列の一般固有値問題に帰着される。なお、この手法ではデータだけでなく、特性項目も同様に1次元軸上の点として数量化される。

4.1 キーワードの選択

この手法では、特性項目をいかに決めるかが重要である。言語テキストの場合、あらかじめ決められたキーワードを含むかどうかを特性項目として定義できる。しかし、キーワード選択の基準はそのほとんどが主観に頼らざるを得ないが、なるべく主観を排除するために、先ほど行った第Ⅳ類による結果を反映させることにした。つまり、第Ⅳ類で得られた空間内で、互いに近いテキストを1つ

のグループとして分類し、そのグループ内で頻度を考慮に入れてキーワードを選択することにした。今回の実験で選択したキーワードは表2の通りである。

あいさつ 受付 英語 駅 大阪 金 支払い
会場 観光 京都駅 公園 小切手 国際会議
記入 希望 質問 住所 用件 参加 参加者
締切り 書類 資料 時間 事務局 組織委員会
地下鉄 聴講 通訳 通訳電話国際会議 手続
手配 東京都 登録用紙 同時通訳 日程 日
本語 番号 費用 申込み 申込用紙 予定 ア
メリカドル クレジットカード スピーチ
ドル バス

表2 キーワード

4.2 実験

実験はADD内のテキスト50ファイルを用いた。結果は、キーワードに関してのみ第Ⅳ類と同様、2次元平面上に配置したものを示す(図3参照)。同じ文脈で用いられる語彙同士が近くに集まっているのが分かる。この数量化された語彙の有無により、テキストを数量化することができる。

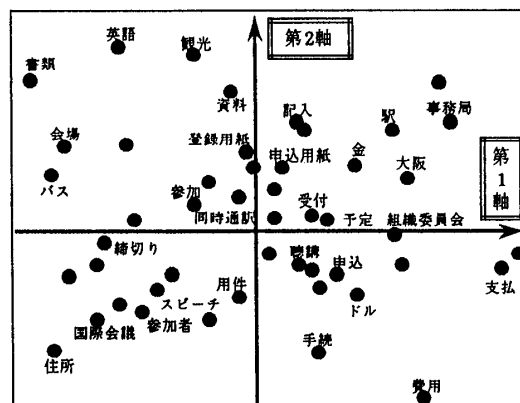


図3 キーワード空間

5. むすび

対話テキストおよび出現するキーワードを2つの数量化法を用いて数量化する実験結果を述べた。同時に、テキストの数量化からもたらされるテキストの分類、具体的には固有方程式を解いて得られる固有軸の意味を基にテキストを分類した。これは、表層上の語彙レベルの分類であったが、定量的なテキスト分析の見通しを得た。今後は、テキスト数を増やし、語彙だけではなく、構文面からのテキストの数量化と分類を試みる予定である。

謝辞

研究の機械を与えて頂いたATR自動翻訳電話研究所 榎村明社長に深謝します。また熱心に討論して頂いた同研究所データ処理研究室の皆様へ感謝致します。

[参考文献]

- [1] 江原:「対話データベースからの統計情報の抽出」情報処理学会 1990年秋期全国大会
- [2] 水谷:「数理言語学」(培風館)
- [3] 西村・岩坪:「計算意味論の実験」情報処理 Vol.11 No.3 Mar.1970
- [4] Biber: "A typology of English texts" Linguistics 27, 1989
- [5] 林:「数量化理論とデータ処理」(朝倉書店)