対話データベースの自動プロファイリング：効率的話題タグ付与をめざして

田中 英輝　　Kristiina Jokinen　　横尾 昭男
ATR 音声翻訳通信研究所
{tanakah, kjokinen, ayokoo}@itl.atr.co.jp

1 はじめに

対話管理の研究においては、発話を特徴づけるさまざまな情報を付与したコーパスが有効である。このなかで従来良く使われてきたのは話者の「命令」や「依頼」といった発話意図の情報である。このような情報を付与したコーパスを作成して発話の管理や予測を研究したものには [4][5] などがある。

一方、最近では発話の意図だけでなく内容を考慮した対話管理の研究が着目されている [1] [1]。著者らも発話の「話題」、すなわち「何に関する発話であるか」に着目した対話管理の研究を進めている [2]。現在はいくつかのタスクにおいて話題の種類を分類してタグの形でコーパスの発話に人手で付与する作業を実施中である。

しかし話題タグの付与には2 節で説明する問題があるため効率的に作業するのは難しい。本稿ではまず話題タグ付与作業の問題を考察して、次にこれをを解決するための手法を提案する。また実際に提案手法を使ったタグの付与実験を行ったのでその結果についても報告する。

2 話題タグ付与の問題

あるタスクに関する対話コーパスがあり、これに話題のタグを付与する状況を考えよう。話題タグを新たにコーパスに付与するには下記の作業が必要となる。

- タグセットの決定
  最初に話題のタグセットを設計しなくてはならない。これにはコーパスを調査してタスクに関連した発話を抽出した上で主要な話題を把握しなくてはならない。この作業はコーパス全体を詳細に調査することになるため労力が大きい。さらに、タスクに関連した発話を抽出しても各発話の話題を決定するのは難しい。話題の決定方法を明示的に示しにくいからである。
  また発話意図の場合は発話行為の研究から続く長い研究の歴史があり、この中で提案されてきたタグセットを参考にすることができる。もちろん対象領域によってタグセットを変更しなくてはならないだろうが、発話意図の種類はそれほど変化しないであろう。しかし、話題は対象領域によって全く異り、また研究の蓄積も少ないため参考にできるタグセットは入手困難である。
- データの品質の確保
  タグを付与したデータはいろいろな使い道がある。この中では、現発話のタグの同定や次発話のタグの予測が重要となる。このためタグを一貫して付与しておくことが必須である。しかし作業者間の揺れや同一作業者でも経時的な揺れを押さえるのは難しい。

3 提案手法

ここでは、上記の問題を解決することを目的としたタグの付与手法について説明する。この手法の骨子は発話集合をクラスタリングしてこれをユーザに提示することで話題タグセットの設計を支援することと、タグのついたデータの一貫性を統計的に評価することにある。

(1) プロファイルの作成とタグセットの設計
  コーパス中の発話を 4 節で説明するアルゴリズムでクラスタリングして階層木を作る。この階層木は4 節で説明するようにコーパス中の主要な話題の構造を示すことができる。このため、話題のタグセットを設計する時にこの情報を参考にすれば手間を軽減することができる。本稿ではこの階層木を話題のプロファイルと呼ぶ。

(2) タグの付与
  次にタグを各発話に付与する。この時同じクラスタに分類された発話は似た話題を扱っている可能性が高いことに注意してタグを付与する。

(3) 一貫性の確認
  タグ付与作業の一貫性を統計的に評価する。タグの利用方法は未知であるため単純な評価手法を採用する。著者らは発話中の単語を使ってタグを決定木で予測した時の精度で評価することにした。ここで十分な予測精度が得られなければ、単語とタグの間の

[1] 発話の「意図」と「内容」という用語は文献 [1] の用語 "intentional" と "informational" に従って使用した。

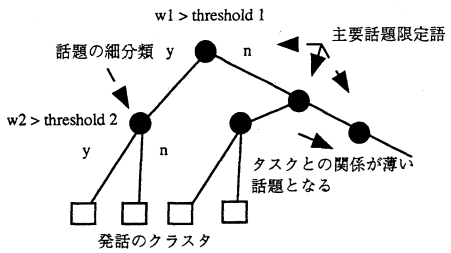

図 1: 階層木とその特徴

相関が低いことになりタグを一貫して付与していない可能性がある。このような場合にはタグ付与を見直すか、タグセットを設計し直してタグ付与作業を実施する。

以上のように (1) でタグセットの設計の手間を軽減し、また (3) で一貫性の確認を行うことで 2 節の問題を解決する。

4 アルゴリズム

本研究で使用するアルゴリズムの説明を行う。本研究で使ったクラスタリングアルゴリズムは、そもそも大規模な文書集合を高速にクラスタリングすることを目的に提案されている [6]。このアルゴリズムは文書の「話題」に関して次のような仮定を設けている。

同一話題の文書集合には「それらのみ」に限って出現する特徴的な単語がある。逆に文書集合にきわだって出現する単語は話題を限定する力を持つ。

著者らは上記の文書集合を発話集合に置き換えた仮定を行ってこのアルゴリズムを適用する。また発話集合の部分集合にきわだって出現する単語を話題限定語と呼ぶ。例えば航空機予約タスクを考えた場合「キャンセル」という単語を一回以上含んだ発話は座席やチケットのキャンセルの話題を扱っている可能性が高い。

このアルゴリズムは「キャンセル」のような話題限定語とその閾値頻度を求め、これを使って逐次二分割クラスタリングを行う。すなわち、最初に発話集合全体を一つのクラスタとする。そして最良の話題限定語とその閾値頻度を求めて、この語を閾値以上含む発話集合とそれ以外に分割する。以後二つの部分集合に対して同様な二分割を再帰的に繰り返す。最良の話題限定語は、発話集合における単語の出現頻度分布を使って求めている。詳細は文献 [6] を参照されたい。

求められるクラスタの階層木は図 1のようになる。このクラスタ階層には以下の特徴がある。

(1) 各ノードには話題限定語とその閾値頻度が対応する。この語を閾値以上含んだ発話は左の辺の下に分類される。各ノードは発話にこのような検査を施す一種の決定木となる。

(2) 階層木の最右辺のノードは話題の混合の大きな発話集合の分割点である。すなわちコーパスの話題を大まかに分類している話題限定語である。これらを主要な話題限定語と呼ぶ。

(3) 左に延びる辺上にある話題限定語は話題の細分類である。

(4) 最右辺上で右にいくほど個別的な発話が分類されている。右に分類されるのは話題限定語を含まない (閾値以下である) という条件に従った発話である。このため右にいくほど主要な話題を含む可能性は低くなる。

階層木が以上の性質を持つため、階層木の左部分に着目することでコーパス中の主要な話題を把握することができる。この階層木のことを話題のプロファイルと呼ぶ。

5 タグ付与実験

5.1 データ

3 節で提案した手法を使って話題のタグを付与する実験を行った。対象としたのは音声対話データベース [3] の航空機予約課題に関する 367 個の日本語発話である。この発話集合に対してまず下記の前処理を施してクラスタリングの入力データである単語の頻度表を作成した。

(1) 発話意図タイプの手がかり表現の削除
発話意図タイプを示す表層表現はすでに収集してあり、これに相当する部分は各発話から削除した。これらは話題には貢献しないと考えたからである。このような表現の大半は「＊していただけますか」や「＊をお願いします」などの文末表現である。

(2) 内容語の抽出
品詞情報を利用して発話中の名詞類（サ変名詞を含む）のみを話題の候補として残した。

(3) 品詞を使った抽象化
品詞名が「番号」「日付」「人名」「会社名」に相当する単語は品詞名で置換した。

上記の前処理の結果 187 の発話が話題の候補単語を含んでおり処理の対象となった。

5.2 プロファイル作成

得られた 187 発話を対象にしてクラスタリングを行いプロファイルを作成した。発話集合の分割はクラスタの大きさが 20 以下で停止するようにした。図 2 に得られたプロファイルの一部を示す。紙面の都合で木を回転して左上が木のルートノードになるように表示している。図中の黒丸が話題限定語を示すノードである。“*”

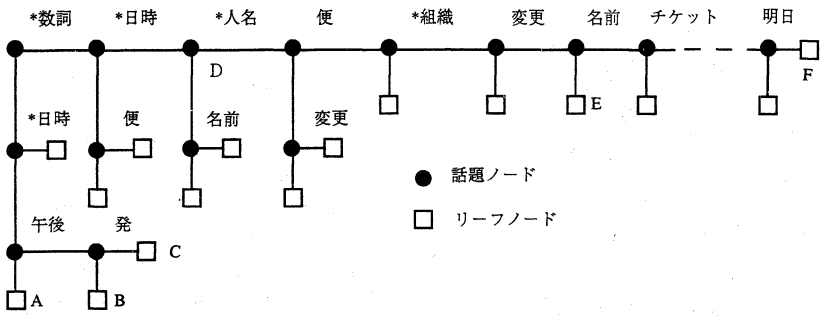

図 2: 話題プロファイルの一部

表 1: 主要な話題限定語の頻度分布

*数詞	*日時	*人名	便
33	21	24	25
*組織	変更	名前	チケット
14	13	8	7
ニューヨーク	予約	アイ	満席
4	5	1	4
客	空席	次	明日
2	2	2	2

で始まる話題限定語は品詞である。その他の話題限定語は表層の単語である。

話題「* 数詞」「* 日付」の閾値頻度は 1.5 でありその他は 0.5 であった。閾値頻度を越えて話題を含む発話は垂直な辺の下に分類されていることに注意されたい。

表 1 に主要な話題限定語とそれを閾値以上含んだ発話の数の分布を示す。

これらのデータによって全発話の話題の特徴を把握できる。ここでは形式的な特徴を示す。表 1 から最初の 4 つの主要な話題限定語だけで全体の発話の 55% (103/187) を含んでいることがわかる。

プロファイルの深い部分の発話は複数の話題限定語を含んでいるため長くなる傾向にある。たとえはリーフ A と B は 4 つの話題限定語を含んだ長い発話である。一方浅い部分は短い発話の可能性が高い。このプロファイルでは右側に分類されている発話ほど短くなっている。また次に説明するようにこの部分は主要な話題を含まない発話が分類される傾向が強い。

5.3 タグ付与と評価

次にこのプロファイルを参考にしてタグセットを作成した。具体的な手順を説明する。例えば、最も重要な話題限定語は「* 数値」である。これは実際にはすべて飛行機の便名に対応した数値表現であり、この下には便名を含んだ発話が分類されている。例えば「* 数値」と「* 日時」を含んだ発話 (リーフ C) には「八月十日、二百一便でございますね。」、「八月十一日、明後日の二百一便ですね。」 、「八月十一日木曜日、二百七便でございます。」などがあった。これらの発話は「便名」が話題である。そこで「flight-num (便名) 」という話題タグを設定した[2]。

「* 人名」 (ノード D) と「名前」 (ノード E) は一見似たような話題限定語であるが、これらのノードに分類された発話の話題はかなり異っていた。ノード D の下には「鈴木さんですね。」、「私、担当の鈴木と申します。」のように話者の名前を告げる発話が分類されている。これらの例より「name (人名) 」というタグを設定した。一方、リーフ E には「ではお名前と日付、便名をお願いいたします。」「お名前をおっしゃってください。」というような顧客の情報を確認する発話が分類されており、これらから「client-stat (顧客の予約状況) 」というタグを設定した。

リーフ F にはタスクに関係の薄い発話が分類されていた。例えば、「それは簡単です。」、「チェックしてみます。」「都合が悪くて。」である。これらには「mixed (タスクと無関係な発話) 」というタグに相当すると判断した。

以上のようにクラスタに分類されている発話を観察してタグを設定して次にこれを各発話に付与していった。作成したタグセットの一覧とその付与頻度を表 2 に示す。

このような形でタグを設定してタグを付与しても同一クラスタの中の発話がすべて同じタグになるとは限らない。しかし、同一クラスタには似たような話題が分類された場合が多い。例えば、4 つのクラスタを含む「* 数値」の下の発話に付与されたタグの分布を表 3 に示す。このように大きなクラスタのまとまりでさえ全体のタグの頻度分布にくらべて片寄った分布をしていることがわかる。小さなクラスタの場合はほぼ同一のタグを付与できた場合が多い。「* 人名」の下の発話はすべて “name” となった。また「* 組織」の下の 14 の発話の

[2]発話の意図としては「確認」「伝達」などが混ざっている。

表 2: 全タグセットとその分布

話題タグ	付与頻度	説明
mixed	31	タスクと無関係な発話
name	24	人名
flight-chan	22	予約変更
client-stat	20	顧客の予約状況
flight-num	19	便名
flight-sch	19	飛行スケジュール
res-stat	16	飛行機の予約状況
ticket-proc	13	切符の手続き
flight-res	11	予約
flight-cancel	7	キャンセル
date	5	日付

表 3: 「* 数詞」の下の発話のタグ分布

話題タグ	付与頻度	説明
flight-num	16	便名
client-stat	7	顧客の予約状況
res-stat	4	飛行機の予約状況
flight-chan	3	予約変更
flight-cancel	2	キャンセル
flight-res	1	予約

うち 12 の発話は “mixed” のタグを付与した。これらはすべて「エアバシフィックでございます。」のように対話の最初の会社名を告げる発話である。またリーフ F の 20 個の発話の 14 個には “mixed” というタグが付与された。4 節で述べたように主要な話題から離れた話題が右方向に分類される傾向があることを裏付ける結果である。

次にタグ付与の一貫性の評価を決定木を使って行った。入力はクラスタリングと同じく各発話に出現した名詞の頻度であり、これを使ってタグを予測する決定木を学習した。正解率の評価には 187 回の交差確認法を利用した。この結果平均 78 % の正解率となった。著者らは発話意図の手がかり語とタグに対して同様な評価を行って 80 % 程度の正解率を得ており今回の正解率はほぼこれに近い結果となった。これは話題タグの設定とその付与に自立語の情報を利用した効果だと考えている。

6 おわりに

発話データベースに話題タグを効率的かつ一貫した品質で付与するための手法を提案した。この手法の主要な部分はコーパスの発話を話題でクラスタリングして話題のプロファイルを作成し、タグの設計とタグの付与を支援する部分である。

著者らは発話意図のタグおよび話題のタグを英語と日本語のコーパスに付与する作業を現在進めている。これまでは支援環境がない状態でこの作業を行ってきたが、今回の実験でタグ付与に有効であるとの結果を得たので今後は実際の作業に応用したいと考えている。

最後に本研究の課題を記す。プロファイルを作成するアルゴリズムは、話題の候補を発話中の単語で近似している。この近似は多くの場合で正しいと考えられるが、成立しない場合がある。話題を含む単語が省略された場合や、文脈がなければ話題を決定できない発話があるからである。これらを決定するには前後の文脈を作業者に提示することが必要となる。タグ付与の支援環境にはこのような機能も必要である。

参考文献

[1] R. A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, and V. Zue. Survey of the state of the art in human language technology. Technical report, CSLU, http://www.cse.ogi.edu/CSLU, 1995.

[2] K. Jokinen and T. Morimoto. Topic information and spoken dialogue systems. In *NLPRS-97*, pp. 429–434. Proceedings of the Natural Language Processing Pacific Rim Symposium 1997, Phuket, Thailand, 1997.

[3] T. Morimoto, N. Uratani, T. Takezawa, O. Furuse, Y. Sobashima, H. Iida, A. Nakamura, Y. Sagisaka, N. Higuchi, and Y. Yamazaki. A speech and language database for speech translation research. In *Proceedings of ICSLP '94*, pp. 1791–1794, 1994.

[4] M. Nagata and T. Morimoto. An information-theoretic model of discourse for next utterance type prediction. *Transactions of Information Processing Society of Japan*, Vol. 35, No. 6, pp. 1050–1061, 1994.

[5] N. Reithinger and E. Maier. Utilizing statistical dialogue act processing in verbmobil. In *Proceedings of the 33rd Annual Meeting of the ACL*, pp. 116–121, 1995.

[6] H. Tanaka. An efficient clustering algorithm based on the topic binder hypothesis. In *NLPRS-97*, pp. 387–392. Proceedings of the Natural Language Processing Pacific Rim Symposium 1997, Phuket, Thailand, 1997.