非タスク指向型対話エージェントの設計と評価に関する研究

磯村直樹

目次

図目次

表目次

第 1 章

序論

1.1 研究の背景

1.1.1 社会的背景

近年，コミュニケーション不足，コミュニケーションの断絶を原因とする様々な現象が顕在化し，大きな社会問題となっている．例えば，厚生労働省が平成 16 年度に発表した「YES-プログラム[*1]」では，コミュニケーション能力を企業が若年者に求める就職基礎能力と定め，重要視している．しかし，コミュニケーション能力不足や，就職してからの人間関係に対する不安のために，就職活動に消極的な人が増えている．このように，コミュニケーション能力不足の問題は，我々の日常生活だけでなく，企業の雇用にも影響を与えている．前述の YES-プログラムでは，コミュニケーション能力を含む就職基礎能力について，「基礎的なものとして比較的短期間の訓練により向上可能な能力」と定めている．しかしながら，コミュニケーション能力を向上させる機会は少なく，系統的な学習法も確立されていない．コミュニケーションスキルは，経験を重ねることにより向上すると思われるが，コミュニケーションの苦手な人は，人とコミュニケーションを行うことそのものを躊躇するため，能力向上の機会を逃し，悪循環に陥ることになる．

コミュニケーション不足の影響は，雇用問題だけに留まらない．社会の高齢化が進むことによる独居老人の増加に伴い，孤独によるストレスや相談相手の不足が大きな社会問題となっている．厚生労働省による平成 20 年度の国民生活基礎調査 [1] は，全 47957 千世帯のうち，9.06% を占める 4352 千世帯が 65 歳以上の単独世帯であると報告している．また，このような独居老人は 4102 千世帯（平成 18 年度），4326 千世帯（平成 19 年度）と年々増加しており，独居老人でも活き活きと暮らせる社会が求められている．そのためには，他者とコミュニケーションを行う機会を増やすことが不可欠である．実際に，市町

[*1] http://www.mhlw.go.jp/general/seido/syokunou/yes/index.html

村では，生活援助員（ライフサポートアドバイザー）や，友愛訪問活動等の施策がとられてきた．しかし，独居老人の数は増加の一途をたどることが予想されており，介護者の不足によるコミュニケーション機会の減少が問題となる．

このように，コミュニケーション機会の不足は深刻な社会問題となっている．このような問題に対して，ペットロボットのような，見た目やしぐさにより癒しを与えるロボットの研究が盛んになりつつある [2]．しかし，話し相手になることを目的とした対話エージェントの研究は少なく，対話エージェント研究の発展の余地は非常に大きい．そのため，コミュニケーション能力を向上させるための話し相手として，また老人に癒しを与えるための話し相手として今後，一層，人間と対話を行うコンピュータ (対話エージェント) の需要が増すものと考えられる．

1.1.2　技術的背景

コミュニケーションには様々な形態があり，コミュニケーションを参加者数で分類すると，1 対 1 のコミュニケーション，1 対多のコミュニケーション，多対多のコミュニケーションが考えられる．コミュニケーションを方法によって分類すると，対面の音声コミュニケーション，電話などの機器を通じた音声コミュニケーション，メールによる時間差のあるテキストコミュニケーション，チャットなどのテキストによるリアルタイムなコミュニケーションなどがある．コミュニケーションをその目的によって分類すると，何らかの達成したいタスクがあって行うタスク指向型コミュニケーション（映画・レストラン・飛行機のチケットの予約・QA システムなど）と，雑談のように，タスクを持たず，単に楽しむための非タスク指向型コミュニケーションに分けられる．

本論文ではこのように明確な目的を持たない対話を非タスク指向型対話と呼ぶことにする．そして，本研究は 1 対 1 のテキストチャット形式で対話を行う非タスク指向型対話エージェントの設計と評価を対象とする．

従来の対話エージェントの研究は，航空機のチケット予約や接客などを目的としたタスク指向型対話エージェントが中心であり，人間の話し相手になり，対話を盛り上げる非タスク指向型対話に関する研究は少なく，未だ発展途上にある．タスク指向型対話エージェントについては，自動応答システムなど実現されてきているが，非タスク指向型対話エージェントは，対話を破綻せずに続けることさえ困難である．そのため，質の高い質問を通じて対話を盛り上げ，対話相手を楽しませるような，人間らしい対話を可能とするレベルに達していないのが現状である．

対話エージェントの設計開発では，性能の評価が必須である．新しい機能を対話エージェントに実装する際，エージェントの機能改善を試みたときも性能の評価が必要である．これらの効果を評価するためには，客観的，定量的な評価指標が必要となる．例え

ば，タスク指向型対話エージェントはタスク達成率，タスク達成時間などの評価指標を使用できる．反面，非タスク指向型対話エージェントでは，明確なタスクを持たないため，タスク達成率，タスク達成時間などの評価指標を使用できない．そのため，対話をタスクに依存しない指標で評価する必要があるが，このような客観的，定量的な評価法は確立されていない．例えば，非タスク指向型対話エージェントの性能を比較評価する大会である，Loebner Competition [*2]では，人間が主観によって複数の対話エージェント（と人間）を順位付けして評価する．

このように，非タスク指向型対話エージェントの評価には，主に被験者の主観が用いられている．対話エージェントに新しい機能を追加し，その効果を確認するためには，機能を追加する度に性能を評価する必要がある．しかしながら，主観評価では，大量のデータを処理することが難しい．したがって，優れた非タスク指向型対話エージェントを実現するためには，客観的・定量的に対話エージェントを評価する手法を確立することが急務である．本課題を解決し，対話エージェントの比較評価を可能にすることにより初めて，非タスク指向型対話エージェント研究の進展が期待できる．

一方，対話エージェントの機能に関する研究では，ドメインを限定したタスク指向型対話を対象とし，対話に必要な基礎技術の研究が進められている．例えば，旅行やレストラン案内，博物館内でツアーガイド，QA システムなどで，発話の意図，発話の意味の解析や，質問に答えるための情報検索技術などの機能が実現されている．タスク指向型対話エージェントの中には，雑談を行う機能が実現されているエージェントも研究されているが，それらは特定の単語のみに対応している，エージェントの応答の一部を人間が操作しているなど，非タスク指向型対話エージェントとしての機能は不完全なものである．対話処理を自動化するためには，話題の焦点の解析，話者情報の利用，話題に対応した適切な発話の生成など，必要な技術的課題は多い．これらの課題を克服することで初めて，非タスク指向型対話エージェントは実現可能となる．

1.2 研究目的

本研究では，非タスク指向型対話エージェントの実現を目指す．対話エージェントの設計と評価を目標とし，１対１のテキストチャットを行う非タスク指向型対話エージェントを研究対象として採りあげる．人間にとって最も自然な対話は音声対話であるが，本研究では問題を言語処理に特化するため，テキスト対話を対象とし，音声認識・合成技術は評価対象に含めないことにする．

本研究の目標とする対話エージェント KELDIC (Ken's Laboratry Dialogue Com-

[*2] http://www.loebner.net/Prizef/loebner-prize.html

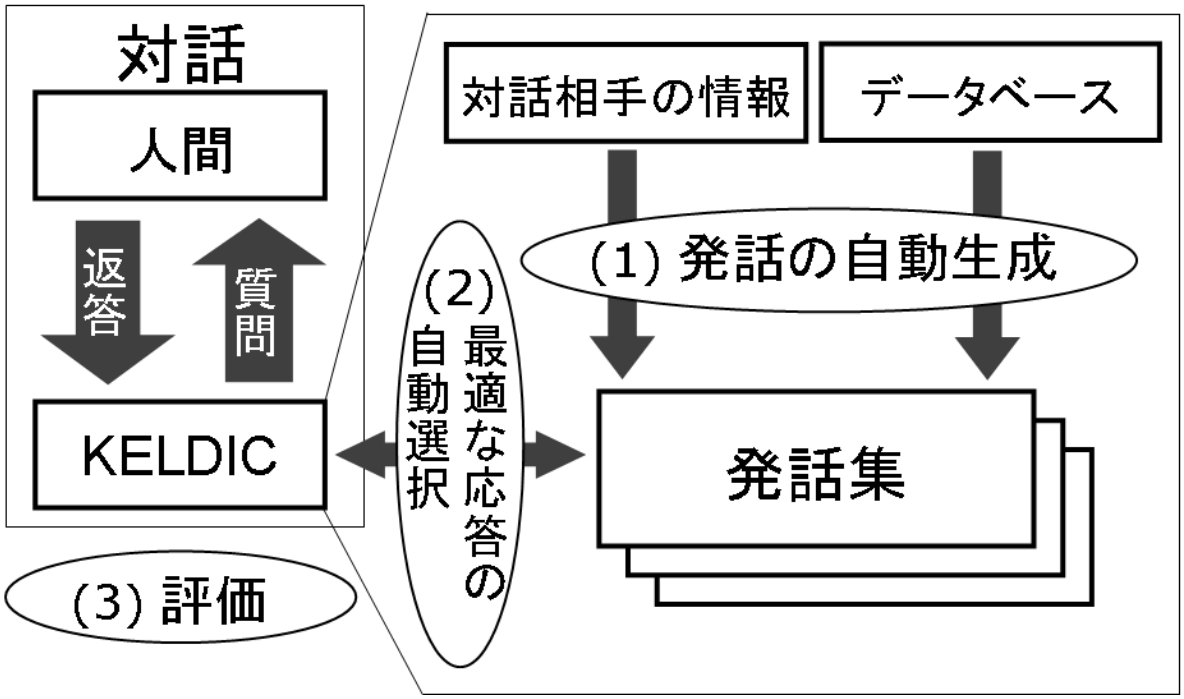

図 1.1　対話エージェント KELDIC の概要

puter) は，質の高い質問を通じて対話を盛り上げ，対話相手を楽しませることを目指している．綿密に練り上げられた質問を用いて対話するためには，対話相手について十分調査し，あらかじめ質問を作成する必要があると考えられる．そこで，綿密に作成された質問のように，あらかじめ発話候補が大量に作成されていると仮定し，その中から発話を選択する手法の確立を目指す．

本研究で目標とする対話エージェント KELDIC を**図 1.1** に示す．

まず (1) で対話相手の情報，Wikipedia や単語 n-gram データベース，これまでの対話内容を基に発話集を生成する．対話相手の情報は，対話相手のブログや著書などから自動的に取得する．特に，趣味や特技などの情報を利用することで，対話を盛り上げる発話を生成可能であると期待される．さらに，Wikipedia などの一般的知識を利用することで，発話の種類を増やし，幅広い話題に対応することが可能となる．

次に，(1) で生成した発話集の中から，(2) で最適な発話を自動選択し，KELDIC の応答とする．最適な発話を自動選択するためには，発話集に含まれる発話を比較評価する手法が必要となる．対話の内容に応じて発話に評価値を付与し，評価値の最も高い発話を最適な発話とすることで，発話集を利用した対話が可能となる．

そして，(3) で対話を評価することで，発話集の自動生成法，自動選択法を改善していく．特に，客観的・定量的な評価は必須な技術であるが，確立されていないのが現状である．

以上のプロセスのうち，本研究では (2) の最適な応答の自動選択と，(3) の評価を対象とする．そこで，本研究の目的を以下のように定める．

【目的 1】非タスク指向型対話エージェントの客観的・定量的な評価法の確立

非タスク指向型対話エージェントを客観的・定量的に評価するための技術を確立する．この技術によって，複数の非タスク指向型対話エージェントの性能の差を数値で確認することも可能となる．

【目的 2】非タスク指向型対話における発話自動選択法の確立

発話集の中から，対話の内容に応じて適切な発話選択を実現する技術を確立する．発話選択法を実現することで，綿密に作成した発話集を利用し適切な返答を返す対話エージェントの実現が可能となる．

1.3 関連研究

1.3.1 対話エージェント

テキスト対話エージェント

テキスト対話を行う非タスク指向型対話エージェントの代表的な例として，ELIZA [3] が挙げられる．ELIZA は入力の表層的なパターンとそれに対する返答の組をあらかじめ用意し対話する非タスク指向型対話エージェントである．ELIZA は発話の意味を理解しているわけではないが，ELIZA と対話した被験者に，人間と対話しているかのように感じさせることができる．これは，発話を理解した上で適切に応答せずとも，人間らしく対話できることを示唆している．

また，意味を理解して対話するタスク指向型エージェントに，SHRDLU [4] がある．仮想世界における円錐，ブロック，四角錐などの基礎的な物体を使用した積み木の問題のみではあるが，これは推論，プラニングといった処理を行うことができる．

本研究で目標とする KELDIC は，SHRDLU とは異なり，ELIZA のように発話の意味の理解，推論，プラニングといった処理は行わない．

音声対話エージェント

音声対話を行うエージェントには 1970 年代に CMU で開発された HEARSAY [5]，MIT の VOYAGER [6] がある．他にも，1990 年代からは，天気情報システム JUPITER [7] や飛行機のチケット予約システム Mercury [8] など，領域を限定したタスク指向型対話の対話エージェントが研究されてきた．また音声対話エージェント開発のためのアーキ

テクチャとして Galaxy[9][10] が提案され，一定の成果を挙げている．マルチモーダル対話エージェントでは，Open Agent Architecture(OAA)[11] を用いて，手書き文字，ジェスチャー，音声を組み合わせた旅行案内エージェントが開発されている [12]．また，近年では，第二言語の教育 (Second Language Acquisition) のための対話エージェントに関する研究事例が Seneff により報告されている [13].

これらのエージェントは全てタスク指向型対話エージェントである．タスク指向型対話エージェントでは，音声認識で扱う単語をタスクに関連する単語のみに制限できるという特徴がある．一方，非タスク指向型対話エージェントでは，対話で扱う単語の数は無数にあり，音声認識は困難である．音声対話エージェントの評価は音声認識の精度に大きく依存するため，音声入力に起因する問題が重要となってくる．また，出力される音声の質（韻律，文の種類）によっても評価が変化する．このように，音声対話エージェントの評価はインタフェースによる影響が大きくなるため，本研究では音声対話エージェントではなく，テキスト対話エージェントを対象とする．

ロボットを用いた対話エージェント

ハードウェアとしてのロボットを用いてタスク指向型対話を行う対話エージェントも開発されてきた．Burgard らは，ナビゲーションスキルを持ち，博物館内でツアーガイドを行うロボットを開発した [14]．Siegwart らは，万博会場内のガイドを行うロボットを開発した [15]．これらは，人とのコミュニケーションや移動を完全自律動作で行い，案内サービスを実現した．ただし，人と対話する際には，特定の単語のみに対応する音声認識機能が用いられており，雑談のような非タスク指向型対話に対応することは難しい．雑談を可能とするため，Wizard of Oz(WOZ) 法を用いた研究もある [16, 17]．これは，ロボットの動作の一部を人間が遠隔操作することで，柔軟な対応を可能にする手法である．しかしながら，WOZ は人手が介入するため，完全に自動化して対話することはできない.

本研究では，最終的にはロボットにすることを目標としているが，前述のように音声認識が困難であるため，ロボットは用いない．そのため，コンピュータ上のプログラムとテキストチャットを行うことで対話を行う.

日本語の対話エージェント

これまで述べてきた対話エージェントの扱う言語は英語であるが，本研究では日本語の対話エージェントを扱う．日本語の対話エージェントでは，バス案内システム [18]，ホテル検索システム [19]，マニュアル検索システム [20] や，MIT の Jupiter を日本語化した Mokusei[21] などがある．日本語の対話エージェントにおいても，英語の対話エージェントと同様，本研究の目指す非タスク指向型対話エージェントは実現されていない.

その他の対話エージェント

タスク指向型対話の一種である，質問応答など情報検索に関する研究は，TREC[*3]，NTCIR[*4]などのワークショップで中心的なテーマとして取り上げられている．NTCIR では，大量の文書データの内容に関する WHY 質問，言語横断質問など，複雑な質問に対する質問応答タスクを実施し，一定の成果を挙げている [22, 23, 24, 25]．また，音声入力による ODQA(Open Domain Question Answering) システムとして，SPIQA[26] も研究されている．このシステムでは，ユーザの音声による断片的かつ曖昧な質問からシステムが必要に応じて確認や問い返しを行い，正しい解答を導く機能を持つ．また，情報を与える形式の一つとして，クイズに着目した対話エージェントの研究がある [27]．これは，歴史上の人物情報などをクイズの形でユーザに提示し，ユーザが学習することを支援するシステムである．この対話エージェントは，ユーザを楽しませるため，対話をクイズというゲーム形式で作成している点が特徴的である．他にも，20 Question のゲーム形式で対話する対話エージェントも研究されている [28]．20 Question は，ユーザが思い浮かべた物事を，20 回以内の yes/no 質問で推定するゲームである．この対話エージェントは，20 Question を通じてユーザから常識を抽出するという点が特徴的である．この研究は，MIT の Open Mind Common Sence プロジェクト[*5]の一環として行われており，このプロジェクトは常識データベースの構築を目的としている．

特に，クイズ対話，20 Question 形式の対話では，対話相手を楽しませるという点において本研究と類似している．ただし，どちらの対話エージェントもタスク指向型対話エージェントであり，本研究の目標とする非タスク指向型対話エージェントとは異なっている．

1.3.2 対話エージェントの機能

発話生成

対話エージェントに必須な機能である，発話生成に関する研究は，テンプレート方式，文生成方式に大きくわけられる．テンプレート方式で発話を生成する代表的な対話エージェントは前述の ELIZA である．また，同じくテンプレート方式の一つである VoiceXML[29] は，音声アプリケーションの構造を統一的に記述するために開発され，W3C 勧告となっている．

テンプレート方式は，テンプレートの作成者が構造を理解しやすいため，テンプレート

[*3] http://trec.nist.gov
[*4] http://research.nii.ac.jp/ntcir
[*5] http://openmind.org/

に対する小規模の追加，修正が容易である．そのため，出力する発話の種類が小規模でよい場合には有効である [30]．しかし，生成すべき発話の種類が大規模になると，テンプレート数が膨大になり，テンプレートの修正が困難になるという問題がある．この問題に対処するため，文生成方式が研究されている．MIT の Galaxy において，返答生成部として使われる GENESIS[31] は，まず，発話を Semantic Frame という意味の構造に変換することで，発話の意図を解釈する．次に，対話管理部で DB を検索するなどして，同様の構造で返答の意味を生成する．そして，発話生成規則に従って意味から発話を生成している．これにより，言語によらず返答を扱うことができ，多言語化が可能となっている．

Stent らは，レストラン推薦システム MATCH(Multimodal Access To City Map) において，学習可能な文プラニング手法である SPaRKy(Sentence Planning with Rhetorical Knowledge)[32][33] を実装している．SPaRKy は，レストランの情報が入力として与えられ，その情報から発話を生成する．1 つの情報から複数の発話候補を生成するため，ランキング処理を行い，1 位にランキングされた文を応答として出力する．ランキングに用いる知識は機械学習によって獲得する．学習では，発話候補を被験者に読ませ，その自然さに対して 5 段階で評価させる．そして，発話候補の持つ特徴を抽出し，RankBoost アルゴリズム [34] によって評価規則集合を学習する．

Oh らは，タグ付与された対話コーパスから，発話生成を行う手法を提案している [35]．この手法では，発話生成に必要な統計モデルを学習するために，旅行予約ドメインの人間対人間の対話コーパスに対して，発話の意味と発話内の内容語の意味をそれぞれ発話クラスタグ，単語クラスタグとして付与している．そして，これらのコーパスを用いて n-gram モデルを作成し，発話生成を行っている．

本研究では，あらかじめ用意した発話集を利用して対話することを想定しているため，テンプレート方式の発話生成であるといえる．ただし，本手法では統計的な手法を用いて発話を選択するため，大規模な発話集に対しても適用可能である．

対話管理

対話管理に関する研究では，対話の構造，主導権，対話制御，ユーザモデルなどの研究が行われている．対話構造の研究では，タスクの構造や，発話者の意図を一般的な形式で表現することで対話の構造を記述するという方法で理論化が行われてきた．計算言語学における先駆的な研究では，Grosz らが談話構造理論を提案している [36]．Grosz らの談話構造理論では，対話を言語構造，意図構造，注意状態という 3 つの概念で構成する．言語構造は複数の談話単位 (Discourse Segment) からなる．談話単位とは，対話を区切り，発話集合としたものである．この言語構造を用いて，意図構造とプランニングを組み合わせることでシステムの意図を生成することが可能である．また，注意状態と発話生成を組み合わせることで応答文生成が可能となる．

対話を状態遷移モデルと捉え，対話を管理する手法も研究・開発されている．有限状態オートマトンを用いる手法では，対話エージェントの発話を状態，ユーザの発話やシステムの処理結果を遷移時の出力シンボルとおいて，システム主導の対話を表現している [37]．このようなオートマトンによる対話制御は，システム開発者にとって対話の流れが容易に把握できるという長所がある．一方で，状態数や遷移条件が複雑になりがちであり，特に，ユーザ側に主導権が移った場合の自由度が高く，保守が困難であるという問題点がある．

このような問題に対処する方法として，マルコフ決定過程を用いる手法がある [38][39]．すなわち，全ての可能な状態および可能な遷移を明示的に表現するのではなく，いくつかの状態を組み合わせ，システム側の意図を示す新たな状態として扱い，対話管理を行う．さらに，このように対話の状態を表現することで，対話エージェントの行動を強化学習の手法で学習することが可能となる．そのため，対話コーパスを用いた対話戦略の自動学習に適している．

タスク指向型対話に適している方法として，知識駆動による対話制御が挙げられる．知識をフレーム [40] や木 [41] を用いて表現し，タスクに必要な知識を表現している．知識を変数として扱い，対話の目標と変数の関係を表現することで，対話を制御する．

本研究では，対話管理は行わないが，非タスク指向型対話エージェントの評価法において対話構造をモデル化している．対話構造のモデル化では，「挨拶」，「質問」といった談話の浅い構造に着目し，状態遷移モデルを作成している．

1.3.3 対話エージェントの評価

タスク指向型対話エージェントの評価

タスク指向型対話エージェントの評価には，タスク達成率や対話の長さ (タスク達成に要した時間) などのコストを用いた客観的・定量的な評価法が確立されている [42][43]．タスク指向型対話エージェントの評価のフレームワークとして知られている，PARADISE [44] は，客観的評価尺度としての対話エージェントの特性と，主観的評価尺度としてのユーザ満足度を組み合わせた評価モデルである．このモデルでは，過去の対話データと人間の主観的評価結果から，ユーザ満足度をタスク達成率と各対話コストの線形結合として表現して，結合の重みは重回帰分析によって求めている．対話コストには応答待ち時間や発話数などが用いられる．タスク達成率はタスクの複雑さによって正規化され，異なるタスクに対する異なる対話エージェントの性能比較も可能である．

非タスク指向型エージェントの評価

タスク指向型対話エージェントとは異なり，本研究で対象とするような非タスク指向型対話エージェントにはタスクが存在しない．そのため，評価にはアンケートなどの主観的な評価法が主に用いられ，対話がうまく行われたかどうかを客観的・定量的に評価する方法はない．

非タスク指向型対話エージェントの評価法として，対話エージェントの応答の意味に基づいた評価法 [45] がある．意味的に正しく表現が自然な応答を正応答，意味的に正しいが表現が不自然な応答を準応答，意味的に誤りである応答を誤応答として，全ての応答に対する正応答と誤応答の割合で評価する．ただし，この評価法では，正応答，準応答，誤応答の判断を人間が経験的に行っている．そのため，この評価法は主観的な評価法であり，大量の対話を評価することは困難である．また，非タスク指向型システムの性能を比較するコンテストとして Loebner Competition が実施されている．これは，被験者がテキスト対話を行い，対話の相手が人間かコンピュータかを評価し，最も人間らしい対話エージェントを選ぶコンテストである．言い換えると，このコンテストはチューリングテスト [46] を行っており，システムが知的であるかどうかを人間が主観によって判定する．2009 年 9 月に開催された大会では，4 人の人間と 3 種類の対話エージェントの合わせて 7 種類の対話を 4 人の人間が順位付けを行った．その結果，対話エージェント “Do Much More” が平均 4.5 位となり，対話エージェントの中で最も良い評価を得ている．

客観的・定量的な非タスク指向型対話エージェント評価法としては，人間同士の対話を基準として，人間同士の対話との類似度という観点から人間らしい自然な対話であるか否かを評価する手法 [47] がある．ただし，この方法は基準となる対話，評価する対話の両方に人手でタグ付与する必要があり，人的コストが大きいという問題がある．

1.4　本論文の構成

以下，本論文の構成について述べる．第 2 章では，非タスク指向型対話エージェントを客観的・定量的に評価する手法について述べる．ここでは，いわゆる対話の「浅い構造」にのみ着目し，発話間の繋がりという最低限の自然さを評価することを試みる．本手法では，人間同士の対話は自然で理想的な対話であると仮定し，人間同士の対話に発話タグを手動で付与し，付与したタグの系列を学習した隠れマルコフモデル (Hidden Markov Model(HMM)) を作成する．人間と対話エージェントの対話を HMM に入力し，その出力確率を人間同士の対話との類似度とみなすことにより，対話エージェントの性能を評価する手法を提案する．実験では，複数の対話エージェントを評価することにより，性能の比較が可能であることを示す．

第 3 章では，第 2 章の評価手法を発展させ，一連の処理のうち，手動タグ付与を自動化する手法を提案する．まず，複数の自動タグ付与の手法を比較評価する．その中で最も精度の高い CRF(Conditional Random Fields) を使い，タグを自動付与する．自動付与したタグの系列を学習した HMM により対話を評価し，手動タグ付与の場合と比較する．これにより，タグ付与の自動化による影響を調査し，評価法としての有用性を確認する．

第 4 章では対話エージェントの発話選択のための，機械学習を用いた統計的手法について述べる．発話選択を，大量の発話候補の中から，入力に対して最も評価の高い発話を選択する問題と捉える．大量の発話候補を順位付けするために，入力と発話候補の対に対して，適切または不適切を手動で評価し，教師データとする．教師データから発話候補の評価値の相対的な大小関係を学習し，発話選択を行う．実験では，発話候補に対する手動評価の結果と自動評価の結果を比較することで，本手法の有効性を評価する．

第 5 章では，本研究で得られた知見，成果をまとめ，対話エージェントの設計と評価に関する今後の研究課題について述べる．

第 2 章

HMM による対話エージェントの評価

2.1 はじめに

対話エージェントを設計するためには，対話の評価尺度が必要になる．タスク指向型対話エージェントの評価には，タスク達成率やタスク達成までに要した発話回数などを用いた客観的・定量的な評価法が提案されている [42, 43, 44]．一方，非タスク指向型対話は，タスク指向型対話と違い，タスク達成率などの明確な評価尺度が存在しない．そのため，対話の品質を客観的・定量的に評価することは難しく，非タスク指向型対話エージェントの評価にはアンケートなどの主観的な評価法が主に用いられているのが現状である．

本章では，非タスク指向型対話エージェントの客観的・定量的評価法を提案する．その際，評価対象とする対話はテキスト対話に限定する．人間にとって最も自然な対話は音声対話であるが，本研究では問題を言語処理に特化するため，音声認識・合成技術は評価対象に含めないことにする．

現状の非タスク指向型対話エージェントは，人間らしい知的な応答をできる段階にないだけでなく，発話間の自然な繋がりという基本的なレベルも実現できていない．そこで，本章では最低限の対話の自然さを比較評価できるよう，発話の繋がりに注目した評価法を提案する．発話の繋がりを表現するため，対話の浅い構造 (shallow discourse structure)，すなわち発話間の隣接を近似できるという特徴がある SWBD-DAMSL(Switchboard Discourse Annotation and Markup System of Labeling) タグ [48] を用いた．この SWBD-DAMSL タグは，話者の目的，発話の焦点や主題といった対話の深い構造（deep discourse structure）は扱わない．本論文でもこのような深い構造は対象としない．

これまでに，人間同士の対話との類似度を算出することにより，非タスク指向型対話エージェントを定量的に評価する方法が提案されている [47]．この方法は，人間同士の対

話を基準とし，人間と対話エージェントとの対話が基準となる対話にどの程度類似しているかを評価しようというものである．類似度が高ければ，その対話は人間らしい自然な対話であるとみなされる．しかしながら，この方法は評価する対話を全ての基準となる対話と比較する必要がある．また，対話と対話の類似度を固定長の発話系列から計算するため，柔軟性に欠けるという問題がある．本章では，テキスト対話に対して手動でタグ付与を行い，付与されたタグ系列を学習した隠れマルコフモデル (Hidden Markov Model(HMM)) の出力確率を計算することにより対話の自然さを評価する方法を提案する．本評価法の最終目標はタグ付与も含め一連の処理を全て自動化することである．ただし，本章では，問題をタグ系列を用いた評価法の妥当性検証に限定するため，タグ付与は手動で行う．これにより，タグ付与の誤りによる評価への影響を排除できる．自動タグ付与については第 3 章で述べる．

以下，2.2 節では，対話の自然さを評価する手法の概要について述べる．さらに，対話を表現するタグ系列である，簡易 DAMSL タグと本研究で使用する対話データについて述べる．2.3 節では，本実験で比較評価する非タスク指向型対話エージェント KELDIC (Ken's Laboratry Dialogue Computer) について述べる．2.4 節では，HMM についての説明と，HMM を使用した対話評価法について述べる．2.5 節では，本手法を用いた非タスク指向型対話エージェントの比較評価の実験結果を示し，それについて考察する．2.6 節では，HMM による対話エージェントの評価についての結果をまとめる．

2.2　対話の自然さ評価法

2.2.1　概要

本節ではテキスト対話を取り上げ，非タスク指向型対話エージェントの評価法を提案する．非タスク指向型対話エージェントの評価尺度としては，対話エージェントの応答の意味的な正しさや，対話の自然さ等，様々な尺度が考えられる．

その中で，本手法では，対話の自然さに着目する．対話の自然さには，論理の一貫性，敬語表現の自然さ，Grice の協調の原則 [49] の遵守などがあるが，ここでは対話の自然さを談話の繋がりの自然さと定義する。人間同士の対話は自然で理想的な対話であると仮定し，人間同士の対話に類似した対話ほど良い評価を与える評価法を提案する．なお，ここでいう対話とは，一回の発言を発話と定義したときに，会話の始まりから終わりまでの発話の系列である．対話エージェントのレベルが未熟である現状では，人間同士の対話か人間と対話エージェントとの対話かを人間の主観で判断することは可能であることから，対話には自然か不自然かを判断するための隠れた構造が存在すると考えられる．

このような発話系列から隠れた構造を学習によってモデル化する手法として，HMM が

広く知られている．入力系列 (時系列) が与えられると，HMM は当該入力系列が生成される確率を出力する．このとき，HMM でモデル化された構造に近い入力系列ほど出力確率が高くなる．この特徴を用いて，HMM の出力確率の大小によってモデルと入力系列の類似性を計算することができる．

そこで，HMM を用いて人間同士の対話をモデル化し，入力された対話が人間同士の対話にどの程度類似しているかを出力確率の大小によって求める．すなわち，出力確率が大きい対話ほど人間同士の対話に類似した対話であるとみなす．

この HMM の学習には，出力シンボルの選択が重要である．本手法では出力シンボルを発話の種類とする．発話の種類を記述するためのタグとして，SWBD-DAMSL タグがある [48]．電話による英語音声対話を収集した Switchboard コーパスを対象としている SWBD-DAMSL タグは，発話の種類を記述するタグであり，48 種類のタグで発話の種類を記述する．タグには，たとえば，Yes/No 質問を表す「qy」や，提案を表す「co」などが存在する (**表 2.1**)．これらのタグの組み合わせを付与することによって発話の意味を表すことができる．

表 2.1: SWBD-DAMSL タグの一部

	タグの種類	内容
1	qy	Yes/No 質問 「ペットを飼ってますか？」
2	qw	WH 質問 「どこのチームが好きですか？」
3	qo	自由回答式の質問 「どう思う？」
4	qyĝ	付加疑問文/確認 「東京にお住まいですよね」
5	co	提案 「お入用ならレシピありますよ」
6	cc	遂行 「私がやりましょう」
7	ny	答え「はい」 「うん」「はい」
8	nn	答え「いいえ」 「いや」「ううん」
9	na	「はい」を使わない肯定

		(「猫を飼っていますか？」) 「3 匹おります」
10	ng	「いいえ」を使わない否定 (「あのドラマを見ましたか？」) 「うちはテレビがないんですよ」
11	no	その他の答え 「わかりません」
12	aa	同意 「ええ」「私もそう思います」
13	ar	拒否 「いいえ」
14	b	相槌 「ふんふん」「なるほど」
15	fe	感嘆 「ええっ」「へえー」
16	fp	開始の言葉 「こんにちは」
17	fc	閉めの言葉 「失礼します」
18	sv	主観的文 「これはおいしいね」
19	sd	客観的文 「これはリンゴです」
20	%	不明語，曖昧な発話

そこで，HMM の出力シンボルとしては SWBD-DAMSL タグを意味的に近いものでまとめた簡易 DAMSL タグを用いる．自然な繋がりを表現するためには，タグの種類は多いほど良いが，本手法ではタグ当りのデータ数を増やすため，タグの種類を減らした簡易 DAMSL タグを用いる．

提案手法における処理の流れを**図 2.1** に示す．

本手法では，発話毎に簡易 DAMSL タグを付与することにより，対話をシンボルの系列で表現する．HMM を用いた処理を行う際には，前処理を行って全ての対話をシンボルの系列に変換する必要がある．タグを付与した対話のうち，人間同士の対話を学習用として HMM によりモデル化する．その後，この HMM に，評価したい対話である人間と対

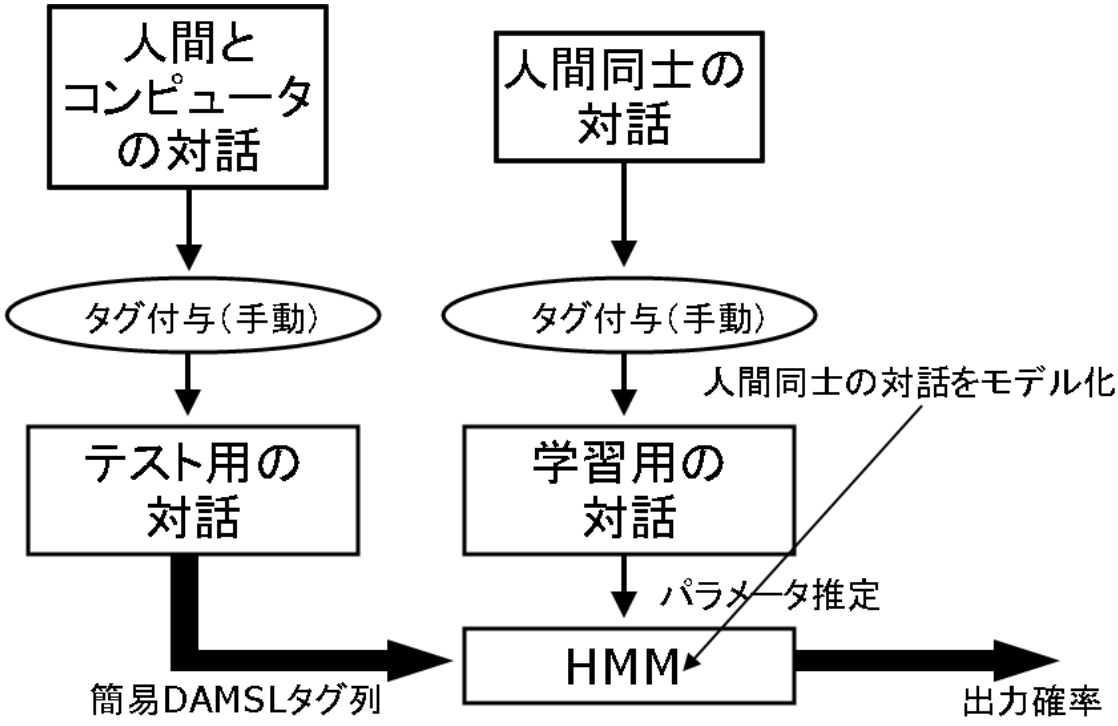

図 2.1 提案手法における処理の流れ

話エージェントとの対話をテスト用として入力する．その出力確率の大小によって，対話の自然さを評価する．

以下が提案手法の大まかな流れである．

1. 学習用発話に手動タグ付与を行う
2. 学習用対話の簡易 DAMSL タグの系列を用いて HMM のパラメータを学習する
3. テスト用の対話 (人間と対話エージェントとの対話) に手動タグ付与を行う
4. テスト用の対話を HMM に入力し，出力確率を求める

以上をまとめると，次のようになる．すなわち，手順 1,2 で人間同士の対話をモデル化した HMM を作成する．その後，手順 3,4 で作成した HMM を用いて評価する．

2.2.2 簡易 DAMSL タグ

本手法では簡易 DAMSL タグを用いる．簡易 DAMSL タグは，SWBD-DAMSL タグをカテゴリ毎にまとめたものであり，**表 2.2** に示した 14 種類のタグが存在する．発話毎に一つまたは複数のタグが付与される．複数のタグが付与される例としては，感謝と説明を同時に行うような発話「Thanking + Statement」や，同意と説明を同時に行う発話

「Agreement + Statement」などがある．未知タグは [Uninterpretable] として表現されている．手動タグ付与の例を**表 2.3** に示す．このように，対話を 14 種類のタグで表現することで，発話間の関係を表現できる．英語文に SWBD-DAMSL タグを手動付与した場合，κ 値が 0.80 と高い値であることがわかっている [48]．なお，κ 値とは複数の付与者の判定の一致率を表す統計量である．本論文では，プロのアノテータが事前に意識合わせを行った上で手動タグ付与を行い，さらに全ての発話において 2 回のチェックを行った．そのため，手動タグ付与のゆれは殆どないと考えられる．

2.2.3 実験で用いる対話

本研究で用いた対話は**表 2.4** の通りである．計 117 対話あり，発話数にすると 8293 発話である．

対話は人間同士の対話 69 対話 (**表 2.4** の (1))，人間と対話エージェントとの対話 48 対話 (**表 2.4** の (2)(3)) より構成されており，対話を行う際に以下の制約を課した．

- 30 分のテキスト対話とする
- 1 対 1 で交互に発話する
- 話題は制限しない
- 顔文字や方言は使用しない

評価対象となる対話の発話数が一定ではないことを考慮し，発話数ではなく 30 分を対話の単位とした．以後，1 対話とは上記の制約を課した 30 分のテキスト対話を示す．これらの制約について被験者に説明した後で対話を収録した．

人間同士の対話は，互いに面識のない被験者 48 人により行われ，全て異なるペアによる対話とした．また，人間と対話エージェントの対話は，それぞれ 24 人の異なる被験者により行われた．

対話エージェントには，著者らが考案した日本語対話エージェントである“ KELDIC(Ken’s Laboratory Dialogue Computer) ” [50] のうち，ELIZA [3] を模して作られた ELIZA 型 KELDIC と，プロのインタビュアを模したインタビュア型 KELDIC の 2 種類を用いた．ELIZA 型 KELDIC はあらかじめ用意された規則の中から，対話相手の発話と一致する規則を探し，定められた返答を行うという対話戦略をとっている．インタビュア型 KELDIC は ELIZA 型 KELDIC を改良した対話エージェントであり，バラエティ番組の司会など，対話のプロであるインタビュアを模した対話エージェントである．インタビュア型 KELDIC は，対話の前に対話相手の情報をもとに，大量の発話集 (スクリプト集) を用意して対話するインタビュー形式の対話エージェントである．

以後の実験では，人間同士の対話 (69 対話) のうち，学習用対話として 45 対話 (**表 2.4**

表 2.2　簡易 DAMSL タグ

	タグの種類	内容
1	Uninterpretable	意味をもたない発話 「あ〜，え〜っと」
2	Self-talk	話者に向かう発話 「何を言おうとしたんだっけ」
3	3rd-party-talk	第 3 者を対象に対する発話 「A さん，はやくこないかな」
4	Statement	思ったこと，説明など
5	Question	質問 「何をしましたか？」
6	Directive	命令 「先にいきなよ」
7	Influencing-addressee-fut-actn	提案 「〜したらどうかな」
8	Committing-speaker-future-action	予定 「〜しないといけない」
9	Other-forward-function	あいさつ 「こんにちは」「さようなら」
10	Thanking	感謝 「ありがとうございます」
11	Apology	謝罪 「ごめんなさい」
12	Agreement	同意 「はい，そうです」
13	Understanding	理解 「なるほど」
14	Other	引用，曖昧な発話

表 2.3　ある対話に対する手動タグ付与の例

発話	付与されたタグ
ありがとうございます	Thanking
アニメに詳しかったりしますか？	Question Statement
あ〜詳しいですよ！！	Agreement
そうなんですかー．どのへんが自分と違う世界だなって思いますか？	Question Understanding Statement

の (1)a) を用い，残り 24 対話 (**表 2.4** の (1)b) はテスト用対話の一部として使用した．対話の種類による差を確認するため，各対話の特徴を比較する．**表 2.4** における，発話数の標準偏差を比較することにより，人間同士の対話に比べて，人間と対話エージェントとの対話は対話毎の発話数の差が大きいことがわかる．形態素数の平均と標準偏差は，発話毎に求めたものである．形態素数の平均を比較することにより，人間同士の対話の方が，人間と対話エージェントとの対話に比べて 1 発話中の形態素数が多いことがわかる．タグの種類数は，全対話に付与されたタグの数である．タグの種類数を比較することにより，インタビュア型 KELDIC と人間との対話はタグの種類数が少ないことがわかる．これは，発話集を生成する際に，数種類の質問生成用テンプレートを使用しているためであると考えられる．

さらに，全ての対話は，対話を行った被験者によって対話の自然さをアンケートにより 1(極めて不自然) から 5(極めて自然) の 5 段階で主観評価されており，その結果は**表 2.4** に示されている．アンケートにおける対話の自然さとは，被験者自身の応答も含めた対話全体の自然さを被験者が主観評価したものである．この結果から，インタビュア型 KELDIC は ELIZA 型 KELDIC より自然な対話が可能であるといえる．また，当然ながら人間同士の対話は最も自然な対話と評価されている．

また，既に述べたように，タグ付与の誤りによる評価への影響を排除するため，全ての対話に対して手動で簡易 DAMSL タグを付与した．上で述べたように，一つの発話に対して複数のタグを付与する場合もあるため，タグの組み合わせを考慮にいれると全ての対話を通してのタグの総数は 89 種類であった．

人間同士の対話で出現頻度が上位 10 位までのタグについて，各対話における相対頻度を**図 2.2** に示す．図より，どの対話も [Statement][Question] が大部分を占めていることがわかる．ただし，[Statement,Question] は人間同士の対話には少なく，人間と対話エー

表 2.4　実験に用いた対話

対話の種類	(1) 人間同士の対話 {48}		(2)ELIZA 型 KELDIC と人間との対話 {24}	(3) インタビュア型 KELDIC と人間との対話 {24}
構成と使用目的	a. 学習用	b. テスト用	テスト用	テスト用
対話数	45	24	24	24
発話数の平均	65.13	66.83	87.00	52.67
発話数の標準偏差	14.17	12.41	52.47	23.25
形態素数の平均	10.55	11.25	8.91	8.58
形態素数の標準偏差	6.71	7.31	5.86	6.00
タグの種類数	54	42	51	23
対話の自然さ	4.12		1.67	2.29

{} 内は被験者数を表す

ジェントとの対話には多い点に差が見られる．これは，対話エージェントが人間の応答を繰り返しつつ質問する特徴があるためと考えられる．各対話の上位 3 種類のタグを比較すると，人間同士の対話，インタビュア型 KELDIC と人間との対話は上位 3 種類のタグで全体の 80% 以上を占めるのに対して，ELIZA 型 KELDIC と人間との対話は 64.6% であり，偏りが少ない．

なお, 全対話を通して最も多いタグの組み合わせは, [Statement, Question]（778 回）である．他には，[Statement, Uninterpretable]（123 回）や，[Statement, Understanding]（67 回）などがタグの組み合わせとして頻出する．

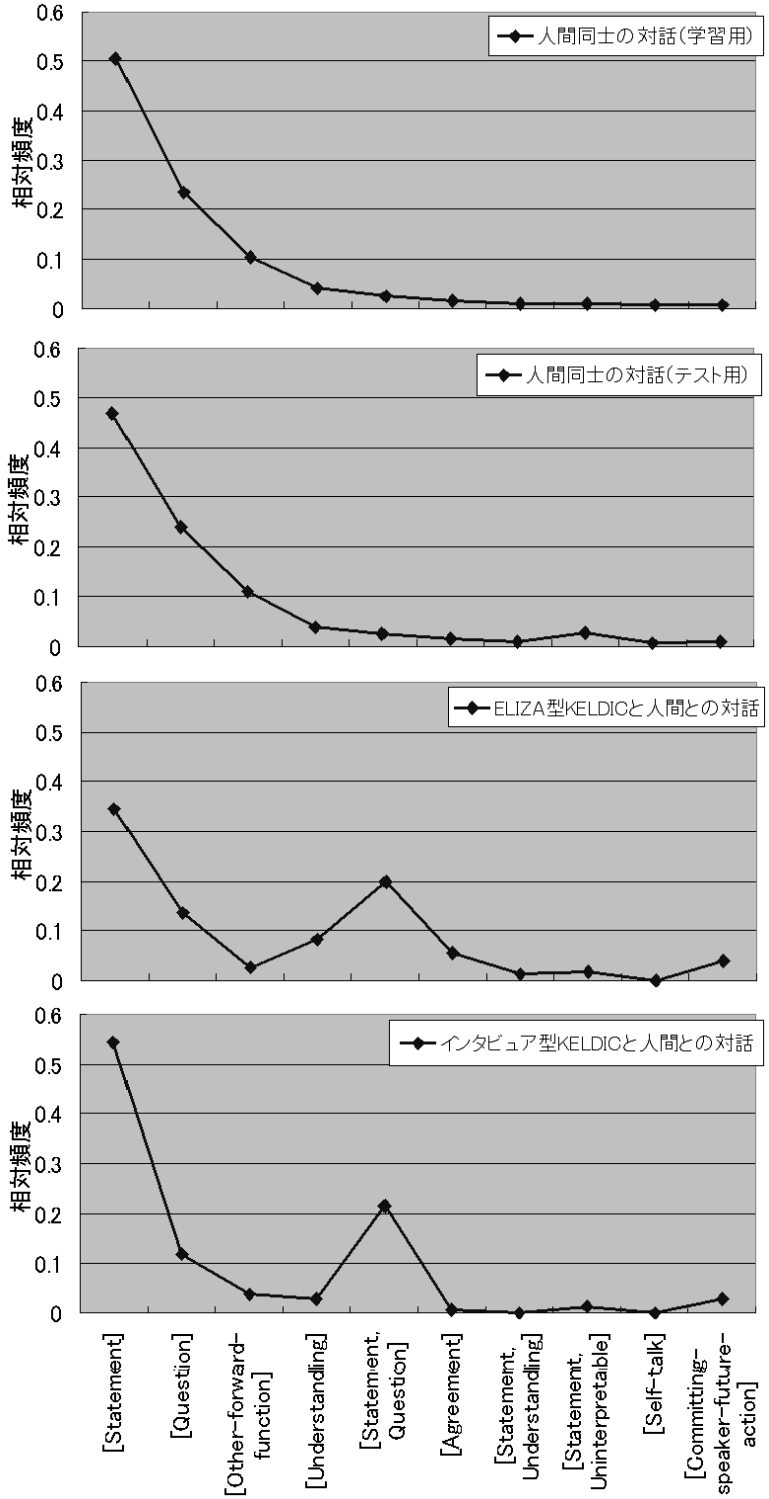

図 2.2　対話毎のタグの出現頻度

表 2.5 ELIZA 型 KELDIC と人間との対話例

発話者	発話
H (人間)	好きな食べ物はありますか？
E(ELIZA 型 KELDIC)	あなたはどう思いますか？
H	味噌カツなどいろいろあります.
E	続けてください.
H	あなたは？
E	自分自身のことですよね？

2.3 実験で用いる対話エージェント

2.3.1 ELIZA 型 KELDIC

対話エージェント「KELDIC」は，著者らが研究している非タスク指向型対話エージェントの総称である．本研究では，非タスク指向型対話エージェントとして，ELIZA 型 KELDIC とインタビュア型 KELDIC の 2 種類を使用した．ELIZA 型 KELDIC の基本的な仕組みは ELIZA [3] と同じであり，対話にはあらかじめ用意された発話生成規則を使用する．ただし，英語ではなく日本語を用いる点が異なっている．ELIZA 型 KELDIC の対話例を**表 2.5** に示す．ELIZA 型 KELDIC は自分のことは述べず，相手の話を聞きだすことによって対話を続ける．例えば**表 2.5** では，人間からの質問に対して意見を述べるのではなく，「あなたはどう思いますか？」のように相手の意見を尋ねるという対話戦略をとっている．

ELIZA 型 KELDIC による発話生成規則

ELIZA 型 KELDIC は対話相手の発話ごとに，対応する発話生成規則を探して返答を生成する．以下に生成規則の例を挙げる．

```
key: 人
  decomp: * (彼|彼女) * 有名 だ 人 * ？
    reasmb: そうかもしれませんね。
    reasmb: そうらしいですね。
```

```
decomp: * 誰 * 有名 だ 人 * ？
  reasmb: あなたは誰か有名人に会ったことがありますか？
  reasmb: あなたは有名人ですか？
```

まず，ユーザの発話を形態素解析し，キーワード (key) を探す．キーワードが見つかった場合，そのキーワードの持つ分解規則 decomp(decomposition) と比較する．形態素解析の解析結果と decomp が一致した場合，再構成規則 reasmb(reassemble) の中から文をランダムに一つ選び，ユーザの発話に対する返答とする．

発話生成規則を用いて，どのような発話に対しても適切な返答を返すためには、広範なキーワードと，キーワードごとに十分な量の decomp，さらに decomp それぞれに適切な reasmb が必要となる．しかしながら，多くの対話に対応するためには膨大な量の生成規則を記述する必要がある．生成規則の全てを人が考え，記述するのは困難であるため，生成規則を自動生成すること求められるが，生成規則は複雑であり，自動生成は難しい．ELIZA は，自分の意見を述べないという対話戦略を採っているため，汎用的な発話が多く，decomp と reasmb の量が少ない場合でも対話が可能である．そのため，本研究では人手で生成規則を作成している．

2.3.2　インタビュア型 KELDIC

インタビュア型 KELDIC は ELIZA 型 KELDIC を改良した対話エージェントであり，バラエティ番組の司会など，対話のプロであるインタビュアを模した対話エージェントである．インタビュア型 KELDIC は，対話の前に対話相手の情報をもとに，大量の発話集 (スクリプト集) を用意して対話を行う対話エージェントであり，インタビュー形式の対話を目指している．一般には，インタビューといえば情報を聞き出す対話であるが，ここでは「徹子の部屋」のような対談番組の司会のように，対話を盛り上げるための対話を目指している．インタビュア型 KELDIC の対話例を**表 2.6** に示す．インタビュア型 KELDIC は対話相手が大学生であるという情報を持っており，「1 番苦労したレポート」についての質問を生成している．また，相手の意見を理解していることを示すために，相手の前の発話を要約して繰り返す戦略をとっている．

インタビュー形式の対話

インタビュー形式の対話では，相手のプロフィールを見て話をリードするインタビュアと，リードされるインタビューイに分かれて対話する．特に，インタビュアは以下の点に注意して対話する．

表 2.6 インタビュア型 KELDIC と人間との対話例

発話者	発話
I(インタビュア型 KELDIC)	大学に入ってから，1 番苦労した レポートは何ですか？
H (人間)	化学基礎実験の無機定性のレポートが 1 番苦戦しました.
I	なるほどね，1 番苦戦したんですね. 化学基礎実験ってどんな感じでしたか？
H	実験の内容はそれほど難しくはないん ですが，課題が難しかったです.
I	なるほど，課題が難しかったんですね. どんな実験を行うんですか？
H	薬品と薬品を混ぜたり分離したりする だけです.

1. YES/NO 型の質問文 (closed question) ではなく 5W1H 型の質問文 (open question) を使う
2. 対話の前にインタビューイについて充分調査する
3. インタビューイの意見に理解を示す
4. インタビューイの好む話題を選択する
5. 安易に話題を転換しない

インタビュア型 KELDIC による発話の生成

インタビュア型 KELDIC はインタビュー形式の対話を行うために，対話の際に 5W1H 型の質問と，相槌を行う．この相槌の処理では，相手の発話を繰り返すことで，インタビューイの意見に対して理解を示している．さらに，対話を続けるために最低限必要となる話題転換，相手の質問への回答も返答とすることで人間と対話する．インタビュア型 KELDIC は以下の手順でユーザの発話から返答を出力する．

1. ユーザが発話をテキストで入力
2. 入力発話に文法的な情報を付与
3. 返答として回答，話題転換，質問，相槌のいずれかを生成
4. 生成された返答を過去に発言していないかチェックを行う．過去に発言していたら

表 2.7　インタビュア型 KELDIC の文法情報

用語	意味
時制	過去・過去進行・現在・現在進行
態	能動態・受動態
体	常体（だ，である調）・敬体（です，ます調）
法	否定・疑問・指示・依頼・意思・忠告・命令・禁止・許可・義務・希望・依頼・推量・確認など

3 に戻って，別の返答を生成

5. 返答を出力

まず，文法的な情報として表 2.7 の時制，態，体，法を付与する．例えば，文節に助動詞「た」を含むとき「過去形」，文節に助動詞の「たい」を含むとき「希望」といったように，複数の規則を設けた．

返答の生成では，入力発話が質問であるかどうかを調べる．入力が質問であれば，KELDIC のプロフィールまたは過去の発話を使い回答する．また，入力発話が否定であればプロフィールに関する質問を生成し，話題転換を行う．入力発話が質問でも否定でもない場合，格フレームを用いて入力発話から質問を生成する [51]．質問を生成できなかった場合，相槌を行う．

2.4　HMM を用いた類似度の計算

本論文では，**図 2.1** のように，HMM によって人間同士の対話をモデル化し，その HMM に評価したい対話を入力することにより対話の自然さを評価する．

本論文で対象とする対話は，始まりの挨拶や終わりの挨拶が存在する系列データであるとみなすことができる．系列データとしては一般に時系列の測定を通して得られる数値データが多い．例えば，ある地域での連続した日々の降水量，外国為替レートの日々の値，音声認識に用いられる音響特徴量などである．また，時系列データ以外では，DNA のヌクレオチド塩基ペア系列などがある．DNA は数値データではないが，ヌクレオチド塩基ペアを種類毎に $1, 2, 3, \cdots$ とシンボル化することで，数値データと同様に系列データとして扱うことができる．

系列データの分析には，マルコフモデルを用いることができる．金融予測などの系列データの応用においては，過去の観測値をもとに時系列における次の値を予測できること

を期待する．将来の値を予測する場合，直近の観測値はそれより過去の観測値よりも有益であると考えられる．マルコフモデルは，この考えをモデル化したものである．すなわち，未来の予測値が直近の観測値以外の過去の観測値に依存しないという仮定を持つモデルである．例えば，マルコフモデルにおける長さ N の観測系列 $x_1, \ldots, x_N$ の同時分布を数式で表現すると以下のようになる．

$$p(x_1, \ldots, x_N) = p(x_1) \prod_{n=2}^{N} p(x_n|x_{n-1}) \tag{2.1}$$

さらに，上式を用いることで，時刻 n までのすべての観測値を与えたときの観測値の条件付き確率は以下のように表される．

$$p(x_n|x_1, \ldots, x_{n-1}) = \frac{p(x_1, \ldots, x_n)}{p(x_1, \ldots, x_{n-1})} \tag{2.2}$$

$$= p(x_n|x_{n-1}) \tag{2.3}$$

これらのモデルの応用では，多くの場合，モデルを定義する条件付分布，$p(x_n|x_{n-1})$ は皆同一であるという制約（定常状態であるという制約）を課す．

しかしながら，対話データにおいては，次の値を予測するときに，直前の観測値以外を含む，いくつかの連続した観測データに影響を受けると考えられる．直前よりさらにさかのぼった過去のデータに影響を持たせるためには，例えば $p(x_n|x_{n-2}, x_{n-1})$ のように高次のマルコフモデルを用いる方法があるが，この方法ではモデルのパラメータ数が増えすぎるため，実用的ではない．

そこで，本研究では高次のマルコフモデルを用いる代わりに，HMM を用いることで対話をモデル化する．HMM は，観測値の出現確率が M 個の状態 (または潜在変数)$S = s_1, \ldots, s_M$ に依存すると仮定し，状態 S がマルコフモデルを構成する，表現力の高いモデルである．HMM において，長さ N の観測系列 $x_1, \ldots, x_N$ の同時分布を数式で表現すると以下のようになる．

$$p(x_1, \ldots, x_N, s_1, \ldots, s_N) = p(s_1) \left[\prod_{n=2}^{N} p(s_n|s_{n-1}) \right] \prod_{n=1}^{N} p(x_n|s_n) \tag{2.4}$$

状態 s_n が観測値 x_n を出力する確率は $p(x_n|s_n)$ で定義される．一般に，HMM は音声認識，自然言語モデル，オンライン手書き文字認識，タンパク質や DNA などの生物学的配列の解析などに広く用いられている．

図で HMM を表すと図 2.3 のようになる．観測シンボル (“Thanking”，“Statement”) が状態 (s_1〜s_4) に依存した確率で表現されるモデルである [52].

例えば，図 2.3 において，現在の状態が s_2 であったとすると，s_2 から s_1 への矢印は，確率 0.2 で s_2 から s_1 に遷移することを示す．同様に，s_2 から確率 0.3 で s_2，0.2 で s_4，0.3

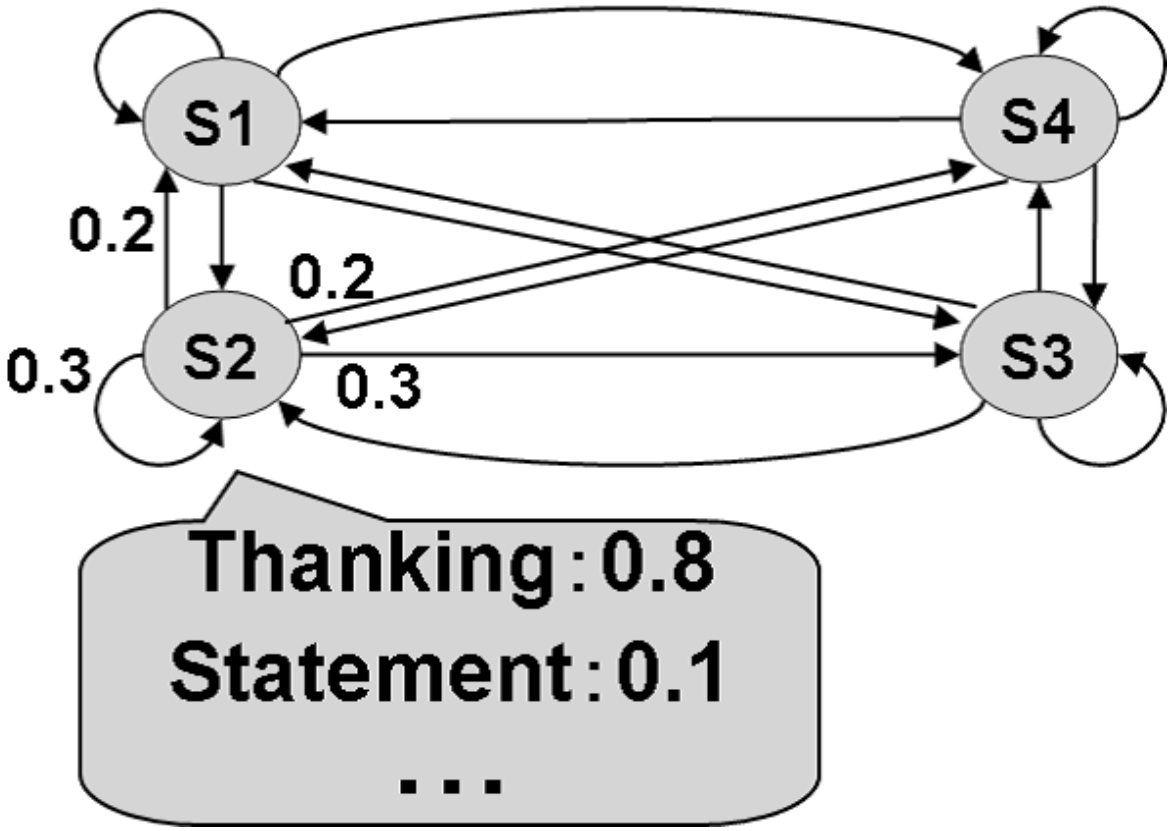

図 2.3　HMM の例

で s_3 にそれぞれ遷移する．また，s_2 に遷移したとき，確率 0.8 でシンボル “Thanking” を出力する．HMM は，全状態 S に対してこれらの遷移確率，出力確率と初期状態を定義することで，シンボル系列の生成プロセスをモデル化している．

HMM はモデルパラメータ λ によって表現される．数学的には，M 個の状態間の遷移確率 $p(s_n|s_{n-1})$ を示す $M \times M$ 行列の状態遷移確率行列 A，K 個のシンボルと状態毎の出力確率 $p(x_n|s_n)$ の分布を示す $M \times K$ 行列の観測シンボル確率分布 B，初期状態を示す初期状態分布 $p(s_1)$ を表すベクトル $\boldsymbol{\pi}$ を用いて，HMM のモデルパラメータ λ を

$$\lambda = (A, B, \boldsymbol{\pi}) \tag{2.5}$$

とする．

モデルパラメータを調整することによって，HMM は様々な観測系列をモデル化することができる．モデルパラメータを調整するために使われる観測系列は学習系列と呼ばれる．学習系列 $X = x_1, \ldots, x_N$ が与えられたとき，HMM のパラメータ λ は，最尤推定により決定することができる．この際，最大化する尤度関数は，

$$p(X|\lambda) = \sum_{S} p(X, S|\lambda) \tag{2.6}$$

である．

モデルパラメータを求めるには，Expectation Maximization(EM) アルゴリズムの一種である，Baum-Welch アルゴリズムを用いる．このアルゴリズムにより，尤度関数を局所的に最大とするモデルパラメータを求めることができる．

モデルパラメータを計算するため，Baum-Welch アルゴリズムでは，最初にモデルパラメータをある初期値に設定する．それをここでは λ^0 とする．そして，$\lambda^{old} = \lambda^0$ とする．また，学習データが与えられているものとする．まず，パラメータの値を用いて $p(S|X, \lambda^{old})$ の事後分布を計算する．次に，この事後分布を用いて，パラメータ集合 λ の関数としての，学習データに対する尤度関数の対数の期待値を求める．この期待値は次式のように関数 $Q(\lambda, \lambda^{old})$ で定義される．

$$Q(\lambda, \lambda^{old}) = \sum_S p(S|X, \lambda^{old}) \ln p(X, S|\lambda) \tag{2.7}$$

この期待値を最大化するように，λ を求める．これは，ラグランジュの未定乗数法を用いることで計算できる．以上の処理を繰り返すことにより，漸近的に尤度関数を最大化することができる．

モデルパラメータを求めることができれば，動的計画法の一種である Viterbi アルゴリズムや forward アルゴリズムを用いて任意の入力系列を生成する確率を求めることができる．本研究では，forward アルゴリズムと Viterbi アルゴリズムの両方で入力系列を生成する確率を計算したところ，差は見られなかったため，Viterbi アルゴリズムを用いて観測系列の生成確率を計算した．

本手法では発話毎に付与された簡易 DAMSL タグを出力シンボルとして用いる．学習系列には人間同士の対話を簡易 DAMSL タグ系列で記述したものを用い，人間同士の対話を反映したモデルを作成する．学習系列が複数あるため，学習系列の各系列毎の尤度関数の総積が局所的最適値になるようにモデルパラメータを求める [53]．学習系列には人間同士の対話を用いているので，出力確率は入力系列，すなわち評価したい対話と人間同士の対話の類似度と考えることができる．

2.5 HMM を用いた対話評価実験

2.5.1 HMM による対話の評価手法

人間同士の対話の学習により HMM を作成し，人間と対話エージェントとの対話を入力したときの出力確率を調べる．また，参考のため人間同士の対話についても評価を行う．実験で使用した HMM の条件は**表 2.8** の通りである．これらの学習用対話を複数入力系列として扱うことで [53]，人間同士の対話を学習した HMM を作成した．以後の実

表 2.8　HMM の条件

状態数	10
出力シンボル	簡易 DAMSL タグの組
学習用対話	人間同士の対話 45 対話 (**表 2.4**(1)a)

験において，状態数による大きな差は見られなかったため，状態数は 10 とした．学習後の HMM の内部構造は**図 2.4** の通りである．図中の楕円は各状態を表しており，Q_0 から Q_9 までの記号で表されている．また，楕円内に書かれている文字は出力シンボルである簡易 DAMSL タグを表す．簡易 DAMSL タグの右隣の数字は，そのタグを出力する確率を表す．簡易 DAMSL タグの種類が多いため，出力確率の大きいもののみを記述している．矢印の太さは遷移確率の大きさを示している．図より，初期状態 Q_0 では挨拶を表す「Other-forward-function」を高確率で出力し，また質問を表す「Question」を出力しやすい状態 Q_2，Q_5 では，思ったことや説明などを表す「Statement」を高確率で出力する状態 Q_3，Q_4 に遷移することがわかる。このことから，人間同士の対話を学習した HMM は，対話の始めに行われる挨拶や，質問に対する返答といった人間同士の対話で見られる構造をモデル化できているといえる．

まず，2.2.3 節で述べた ELIZA を模して作られた ELIZA 型 KELDIC と，プロのインタビュアを模したインタビュア型 KELDIC を比較評価する．使用する対話は**表 2.4** の通りである．

次に，HMM の出力確率を比較することにより，人間の主観と同様に自然な対話，不自然な対話の評価が可能かどうかを調べた．ただし，出力確率は対話の長さで正規化した値を用いた．

また，HMM を用いた評価法の妥当性を確認するため，単純かつ素朴な方法である，学習用対話におけるタグ系列の出現確率に基づく評価法との比較を行った．タグ系列の出現確率に基づく評価法とは以下のような処理である．ここでは，**図 2.5** の対話を例にとって説明する．図中の連続する 2 発話「ありがとうございます」「アニメに詳しかったりしますか？」に付与されているタグである「Thanking」,「Question，Statement」や,「アニメに詳しかったりしますか？」「あ〜詳しいですよ！！」の「Question，Statement」,「Agreement」の学習用対話において連続して出現する確率の平均を評価値とした．すなわち，$(0.7 + 0.1 + 0.3 + \cdots)$/(対話中の連続する 2 発話の総数) がタグ系列のバイグラムによる評価である．この評価値は，学習用対話のタグ系列が評価したい対話にも現れる確率であり，値が大きいほど学習用対話に近いタグ系列であることを示す．

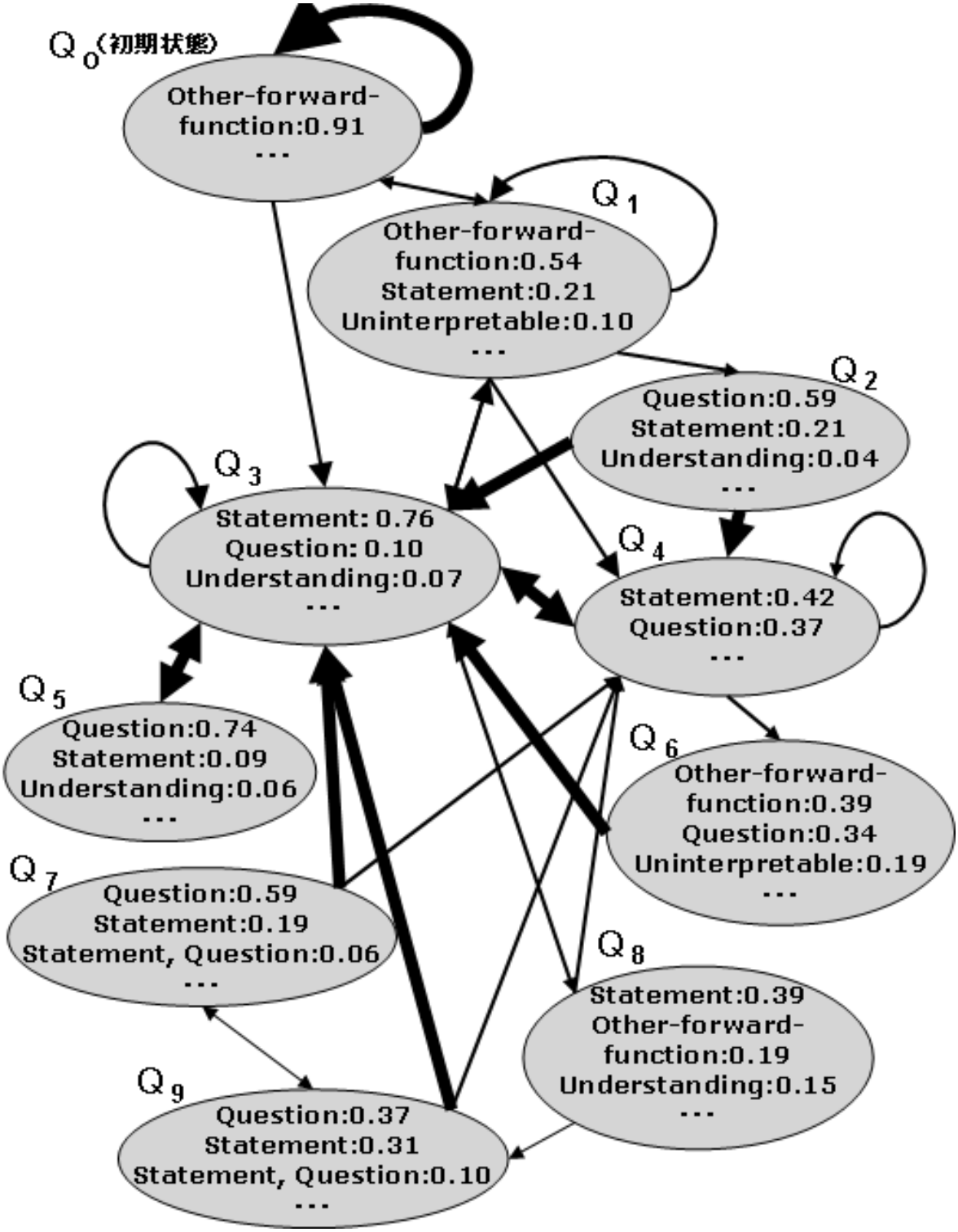
Q 0(初期状態)
Other-forward-function:0.91
...
Q 1
Other-forward-function:0.54
Statement:0.21
Uninterpretable:0.10
...
Q 2
Question:0.59
Statement:0.21
Understanding:0.04
...
Q 3
Statement: 0.76
Question: 0.10
Understanding:0.07
...
Q 4
Statement:0.42
Question:0.37
...
Q 5
Question:0.74
Statement:0.09
Understanding:0.06
...
Q 6
Other-forward-function:0.39
Question:0.34
Uninterpretable:0.19
...
Q 7
Question:0.59
Statement:0.19
Statement, Question:0.06
...
Q 8
Statement:0.39
Other-forward-function:0.19
Understanding:0.15
...
Q 9
Question:0.37
Statement:0.31
Statement, Question:0.10
...

図 2.4 HMM の構造

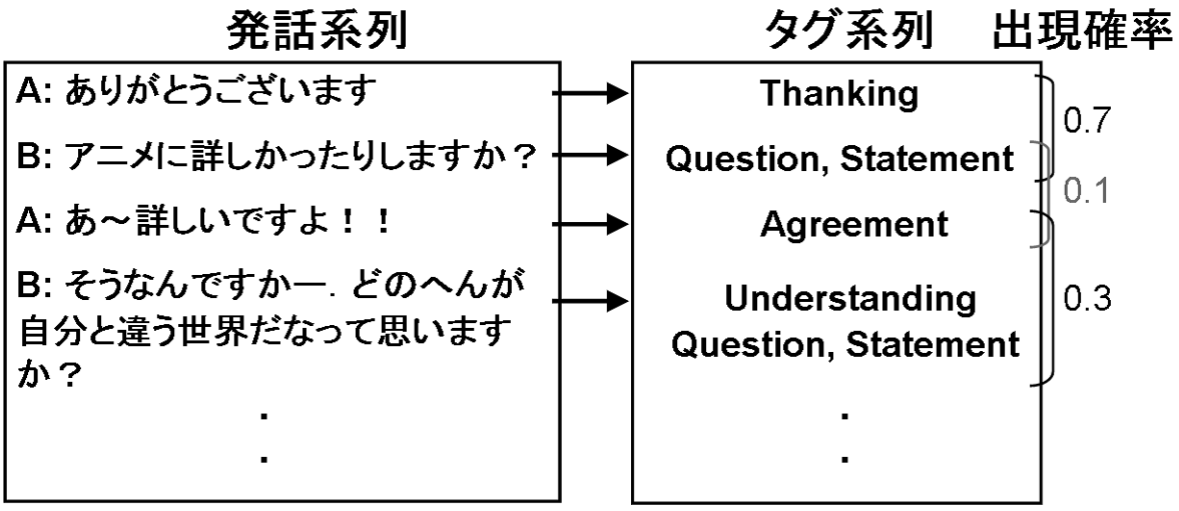

図 2.5　タグ系列のバイグラムによる評価

2.5.2　実験結果

まず，対話エージェントの比較評価の結果を示す．

図 2.6 は (1)ELIZA 型 KELDIC と人間との対話 (**表 2.4**(2))，(2) インタビュア型 KELDIC と人間との対話 (**表 2.4**(3))，(3) 人間同士の対話 (**表 2.4**(1)b) をそれぞれ HMM に入力したときの評価結果である．横軸は HMM による出力確率を表し，縦軸はその相対頻度を表す．同様に，**図 2.7** は**図 2.6** と同じデータをタグ系列の出現確率，すなわちバイグラムによって評価した結果である．**図 2.7** は，前述したように提案手法の妥当性確認のために実施した結果である．

図 2.6 から明らかなように，HMM による評価値は，(1)(2)(3) の順に高くなっており，**表 2.4** と同様の傾向を示していることから，HMM による評価法が妥当であるといえる．一方，**図 2.7** から，単純にタグの出現確率を比較するだけでは対話エージェントの性能を比較することができないことがわかった．なお，ランダムに生成したタグ系列を HMM に入力した場合の出力確率は約 0.002 となり，ELIZA 型 KELDIC と人間との対話と比較しても 10 分の 1 程度の低い値となる．このことからも，HMM による評価は妥当であるといえる．

図 2.6，**図 2.7** は個々の対話を評価し，評価値の相対頻度を示したものである．これらのグラフを視覚的に比較することによっても対話エージェントをある程度評価できるが，より定量的な評価結果を得るため，以下の方法を採る．例えば**図 2.6** の例で，対話エージェント ELIZA 型 KELDIC を評価する場合には，

- ELIZA 型 KELDIC と人間との対話の評価値頻度分布 (**図 2.6**(1)) と,
- 人間同士の対話の評価値頻度分布 (**図 2.6**(3))

との距離を求めればよい．すなわち，この距離が短い程，当該対話エージェントは優れていると判断できる．二つの分布間の距離を計算する方法として，次式の J を用いる．

$$J = \sigma_B^2/\sigma_W^2 \tag{2.8}$$

上式はパターン認識で用いられるクラス内分散・クラス間分散比 [54] に相当し，この値が小さいほど分布間の距離は小さい．ここで σ_W^2 はクラス内分散，σ_B^2 はクラス間分散を表す．この例では，**図 2.6**(1)，**図 2.6**(3) で表される二種類の対話がそれぞれクラスに対応している．これら二つの分布間の距離を J_{13} とおくと,

$$J_{13} = 4.59 \tag{2.9}$$

と求められる．同様にして

- インタビュア型 KELDIC と人間との対話の評価値頻度分布 (**図 2.6**(2)) と,
- 人間同士の対話の評価値頻度分布 (**図 2.6**(3))
  との距離 J_{23} は

$$J_{23} = 0.14 \tag{2.10}$$
$$< J_{13} \tag{2.11}$$

となり，インタビュア型 KELDIC の方が，より自然な対話を実現していることが定量的にも確かめられる．**図 2.7** についても同様に分布間の距離を計算できる．**表 2.9** は，J_{12} や，**図 2.7** の (4)，(5)，(6) から求めた J_{45}，J_{46}，J_{56} を，上記結果と合わせてまとめたものである．J_{13}，J_{23}，J_{46}，J_{56} は理想的な対話 (人間同士の対話) からの距離を示し，値が小さい程良い対話であるといえる．一方 J_{12}，J_{45} は評価対象となる対話エージェント同士の距離を示す．主観評価により対話エージェントの性能に大きな差が見られるならば，J_{12}，J_{45} の値も大きいことが望ましい．

表 2.4 の主観評価と比較すると，HMM による評価である J_{12}，J_{13}，J_{23} は同じ傾向を示している．一方，J_{45}，J_{46} は $J_{45} > J_{46}$ となり，主観評価と一致しない．これは，HMM が任意の長さのタグ系列を高精度にモデル化できたのに対し，タグ系列の出現確率による評価法ではそれができなかったためであると考えられる．これらの評価値の中で，特に重要な値は J_{12} である．J_{12} により，二つの対話エージェントの性能差を定量的に確かめられる．

この結果より，タグの出現確率（バイグラム）を用いた評価法では対話の自然さを評価することができないことがわかる．一方，人間同士の対話を学習した HMM は自然な対

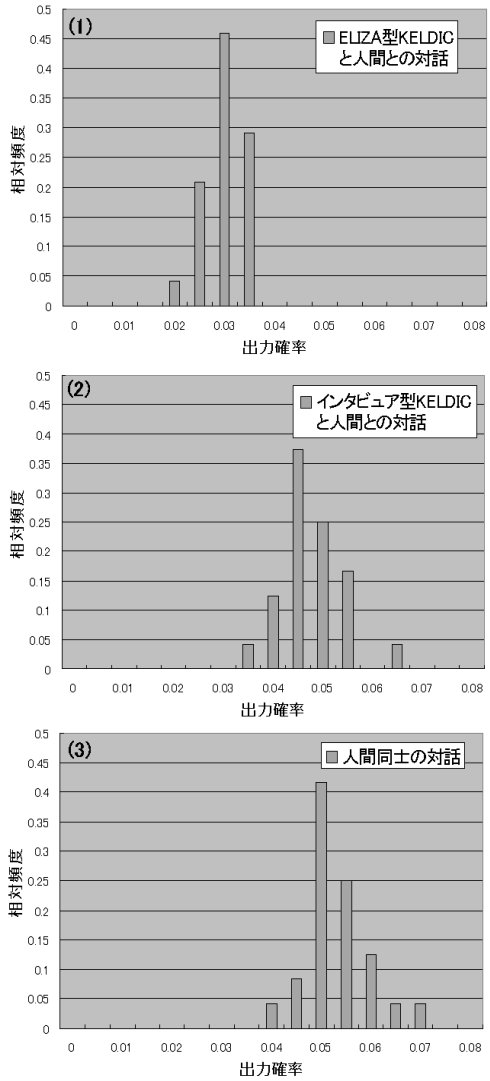

図 2.6　HMM による対話エージェントの性能比較

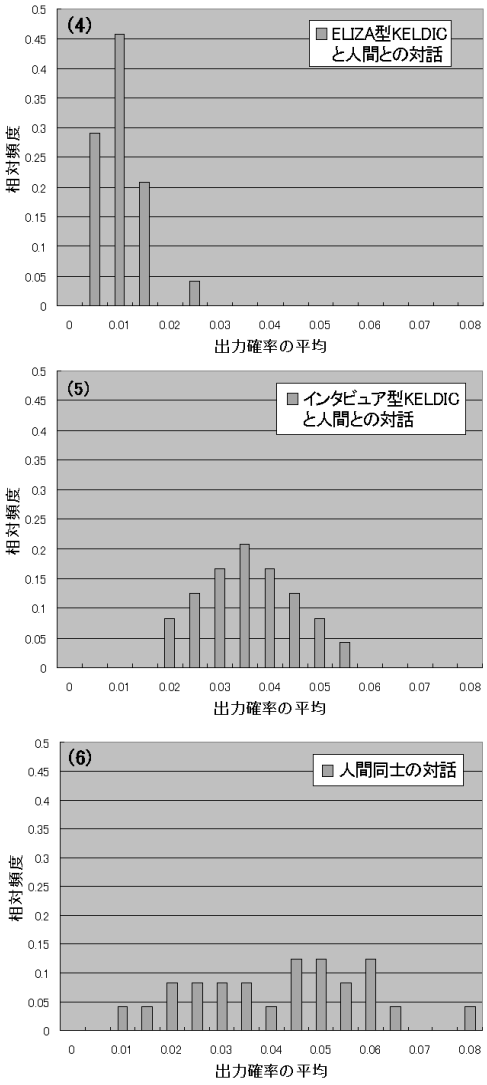

図 2.7　タグ系列の出現確率（バイグラム）による対話エージェントの性能比較

話をモデル化できているといえる．**図 2.6** から明らかなように，現状では人間と対話エージェントとの対話は，出力確率が小さく，不自然な対話であるといえる．人間同士の対話であれば頻度の高い簡易 DAMSL タグの並びが多くなる．例えば，「Question」タグが付与された発話の次の発話は「Statement」タグが付与されることが多かった．一方，人間と対話エージェントとの対話では，対話エージェントの不自然な発話に対して人間が聞き返すことが多くなるなど，人間同士の対話ではあまり現れないタグの並びが多い．そのため，人間同士の対話を学習することにより HMM を作成し，その出力確率の大小を求めることで，対話の評価が可能であることが確認できた．

人間による自然さの主観評価と HMM による評価の比較を**図 2.8** に示す．グラフ上

表 2.9 クラス内分散・クラス間分散比

記号	比較する対話の種類 (数字は**図 2.6, 2.7** のグラフ内の番号を表す)	J_σ
J_{12}	(1) と (2)	2.95
J_{13}	(1) と (3)	4.59
J_{23}	(2) と (3)	0.14
J_{45}	(4) と (5)	2.69
J_{46}	(4) と (6)	1.52
J_{56}	(5) と (6)	0.04

の各点は 1 つの対話 (**表 2.4**(1)b, (2), (3)) に対応している．相関係数は 0.60 であり，HMM による評価と人間による主観評価との間に比較的高い相関がみられた．

表 2.10 は，人間と対話エージェントとの対話の中で，最も大きい出力確率となった対話，すなわち最も人間らしいと評価された対話の一部である．対話の内容を見ても，それほど不自然ではない．

表 2.11 は，人間同士の対話の中で，最も出力確率の小さかった対話の一部である．対話を見ると，発話者 D の発話に付与されているタグが複数のタグの組み合わせになっていることがわかる．複数のタグを組み合わせた発話は学習用対話で出現回数が少ないため，出力確率が低くなったと考えられる．本手法では，簡易 DAMSL タグの並びとして対話を解釈するため，簡易 DAMSL タグの並びが HMM の学習用対話から乖離すれば出力確率にも大きな影響を与える．よって，

- 出現回数の少ないタグの組み合わせが現れた場合
- タグの並びが人間同士の対話と比較して不自然な場合

出力確率が小さくなる．これは，より多くの学習用データを用意することによって解決することができると考えられる．

2.6 要約

本章では，非タスク指向型対話エージェントの客観的・定量的な評価法を提案した．本手法では，対話を評価するにあたり，いわゆる対話の「浅い構造」にのみ着目し，発話間

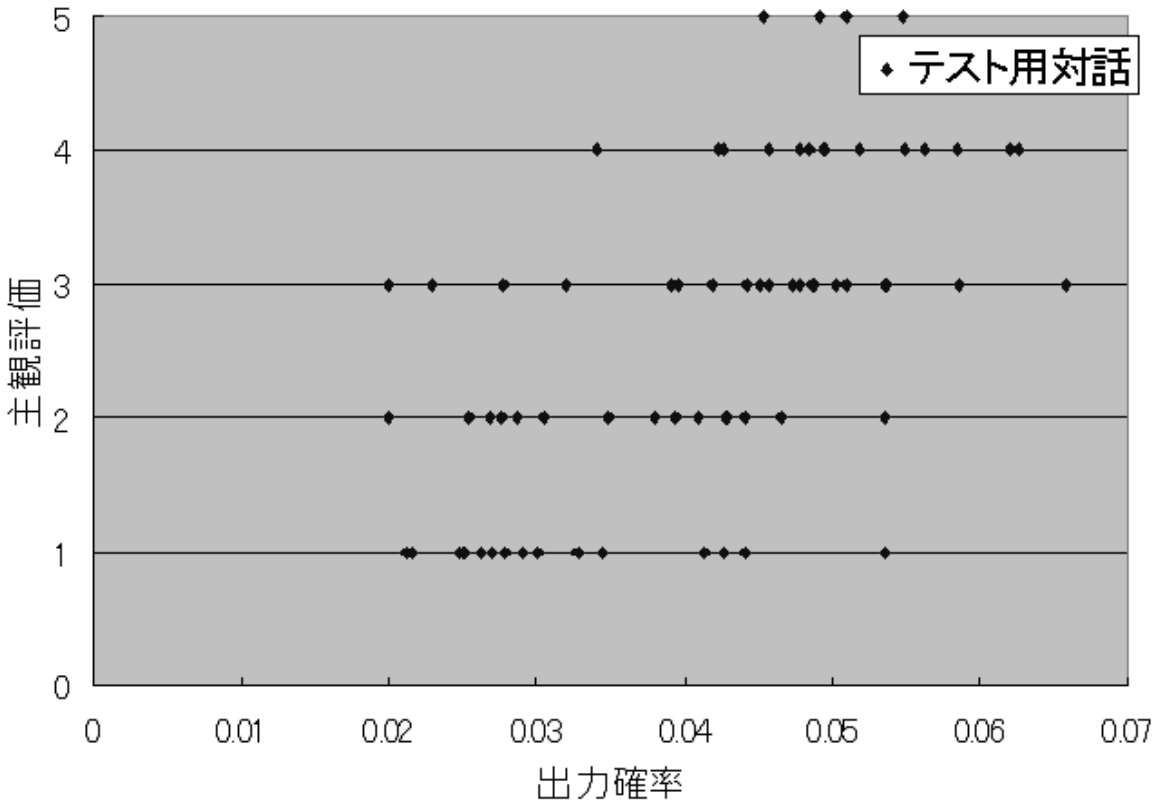

図 2.8　HMM による評価と主観評価との比較

の繋がりという最低限の自然さを評価した．前提として，人間同士の対話は自然で理想的な対話であると仮定し，人間同士の対話との類似度を計算した．

まず，対話をタグ系列として表現した．そして，人間同士の対話タグ系列を学習した HMM を作成した．評価すべき対話をこの HMM に入力し，HMM の出力確率を計算することにより，対話の自然さを評価した．実験の結果，HMM を用いた自然さの評価では，より人間に近い自然な対話を行う対話エージェントほど出力確率が高くなることを確認した．さらに，人間同士の対話は人間と対話エージェントとの対話より出力確率が高くなり，その中でも自然な対話ほど出力確率が高くなることを確認した．これによって，HMM の出力確率の大小によって対話の自然さを評価できることが明らかになった．

本手法は，非タスク指向型対話エージェントを評価の対象としているが，タスク指向型対話エージェントを評価することも可能である．これまでのタスク指向型対話エージェントの評価法では，タスクの達成度や達成時間という側面を評価してきたが，本手法を用いることにより，対話の自然さという側面を評価することができる．

今後，非タスク指向型対話エージェントの性能が向上した場合には，本論文で提案したような，発話間の繋がりの自然さを評価するだけでは不十分である．話者の意図や目的，

表 2.10 人間と対話エージェントとの対話のうち最も大きい出力確率の対話

発話者	発話 [タグ]
A (人)	好きなものはありますか？ [Question,Statement]
B (対話エージェント)	よくわかりません. [Statement]
A	好きな食べ物はありますか？ [Question,Statement]
B	あなたはどう思いますか？ [Question,Statement]
A	味噌カツなどいろいろあります. [Statement]
B	そうそう. [Uninterpretable]
A	あなたは？ [Uninterpretable]
B	自分自身のことですよね？ [Question,Statement]
A	そうですよ [Statement]

発話の焦点，主題等の「深い構造」に着目した評価法に発展させる必要がある．

また，他の側面からの評価も必要である．たとえば，人間を楽しませることを目的とした対話エージェントを設計する際には，自然さだけでなく，対話の楽しさも評価する必要がある．このような場合でも，適切な入力系列と HMM を用いることによって評価することが可能である．

今回の評価は，HMM の学習データとして使用した人間同士の対話との類似度に基づいている．そのため，評価値は相対的であり，評価の絶対値に意味のある，本質的な評価基準も必要である．例えば，2.2 節において対話の自然さの例として挙げた，論理の一貫性，敬語表現の自然さ，Grice の協調の原則の遵守などを評価値とすることが考えられる．これらの値を求める一般的な方法はないが，対話毎に求めることができれば，本質的な評価とすることが可能になると考えられる．

表 2.11　人間同士の対話のうち最も小さい出力確率の対話

発話者	発話 [タグ]
C (人)	駅からちょっと行ったところに, 有名なたこ焼き屋さんがありますよ. [Statement]
D (人)	たこ焼き屋さんですか！ これまた魅力的ですね！ [Understanding,Other-forward-function]
C	駅からの南北はちょっとわかりませんが, 大通り沿いにある大きな服屋さんの裏にあります. [Statement]
D	インターネットで探してみます！ お店の名前ご存じですか？ [Committing-speaker-future-action,Question]
C	ごめんなさい、ちょっと把握してないです. [Apology,Statement]

本章の最初に述べたように，本評価法の最終目標はタグ付与も含め一連の処理を全て自動化することである．ただし，本章では，問題をタグ系列を用いた評価法の妥当性検証に限定するため，タグ付与は手動で行った．手動によるタグ付与は，意味的に正確ではあるものの，手間がかかり，大量の発話を処理することは難しい．そのため，n-gram や SVM を組み合わせるなどの方法で自動タグ付与を行う必要がある．

次章では，自動タグ付与に関わる研究について述べる．

第 3 章

タグ付与の自動化

3.1 はじめに

前章では，テキスト対話を対象とし，Hidden Markov Model(HMM) により非タスク指向型対話エージェントを定量的に評価する方法を提案した．評価対象とする対話をテキスト対話に限定したのは，問題を言語処理に特化するためである．この方法は，人間同士の対話に対してタグを付与し，付与されたタグ系列を学習した HMM の出力確率を計算することにより対話の自然さを評価しようというものである．出力確率が高ければ，その対話は人間らしい自然な対話であるとみなされる．実験の結果，HMM の出力確率の大小によって対話の自然さを評価できることが明らかになった．前章では，タグ系列を用いた評価法の妥当性検証に問題を限定するため，タグ付与は全て手動で行った．しかし，手動タグ付与では大量のデータを扱うことは困難であり，評価法としての有用性を高めるためには，一連の処理を全て自動化することが望ましい．

そこで，本章では，前章の課題であるタグ付与の自動化法を提案する．まず，一定量の対話に手動タグ付与を行い，標準データとする．標準データを用いてテキスト対話に対して自動タグ付与を行う．自動タグ付与の手法は，5 種類の手法を比較評価し，最も高い精度でタグ付与が可能な手法を用いる．その後，付与されたタグ系列を学習した HMM の出力確率を計算することにより対話の自然さを評価し，手動タグ付与を用いた場合と比較することで自動タグ付与の有効性を確認する．

以下，3.2 節では，自動タグ付与を用いた非タスク指向型対話エージェントの評価法について述べる．3.3 節では，自動タグ付与の手法を詳細に述べる．3.4 節では，3.3 節で説明した自動タグ付与手法を比較評価し，実験結果について考察する．3.5 節では，自動タグ付与を用いた非タスク指向型対話エージェントの評価法の有効性を検証する評価実験について述べ，実験結果について考察する．3.6 節では，対話エージェント評価のためのタグ付与の自動化の成果についてまとめる．

3.2　自動タグ付与を用いた対話の自然さ評価法

3.2.1　概要

本節では 2.2 節で述べた対話の自然さ評価法において，タグ付与を自動化した評価法について述べる．タグ付与以外の処理に関しては前章と同様である．すなわち，人間とエージェントのインタラクションのうち，テキスト対話を取り上げ，対話を評価する尺度としては，ここでは談話の繋がりを対話の自然さと定義する．談話の繋がりを表現するため，対話の浅い構造（shallow discourse structure）を近似できるという特徴がある SWBD-DAMSL タグ [48] を用いた．

人間同士の対話を HMM によりモデル化し，評価したい対話を HMM に入力する．そして，人間同士の対話との類似度を HMM の出力確率の大小により求める．すなわち，出力確率が大きい対話ほど人間同士の対話に類似した対話であるとみなす．HMM の出力シンボルとしては，発話に付与するタグである SWBD-DAMSL タグを意味的に近いものでまとめた簡易 DAMSL タグを用いる．

HMM による対話評価法における処理の流れを**図 3.1** に示す．

本手法では，発話毎に簡易 DAMSL タグを付与することにより，対話をシンボルの系列で表現する．簡易 DAMSL タグと HMM を組み合わせることにより対話の評価が可能なことは既に前章で示した通りである．自動タグ付与については **3.3** 節で詳しく述べることにし，ここではまず全体の評価法について説明する．まず，一定量の対話に手動タグ付与を行い，標準データとする．次に，標準データを用いて人間同士の対話と，人間と対話エージェントの対話に対してそれぞれ自動タグ付与を行う．自動タグ付与を施した対話のうち，人間同士の対話を学習用として HMM によりモデル化する．その後，この HMM に，評価したい対話をテスト用として入力する．その出力確率の大小によって，対話の自然さを評価する．

3.2.2　実験で用いる対話

本実験で用いた対話は**表 3.1** の通りである．前章のデータに，標準データを 59 対話 (**表 3.1** の (1)c) を加えた．標準データに対しては手動で簡易 DAMSL タグを付与し，それ以外の対話に対しては自動でタグを付与した．手動タグ付与の例を**表 3.2** に示す．このように，対話を 14 種類のタグで表現することで，発話間の関係を表現できる．なお，未知タグは [Uninterpretable] で表現される．これらのタグは複数付与されることもあり，たとえば，「感謝」と「説明」を同時に行うような発話「Thanking + Statement」や，「同意」と「説明」を同時に行う発話「Agreement + Statement」などがある．

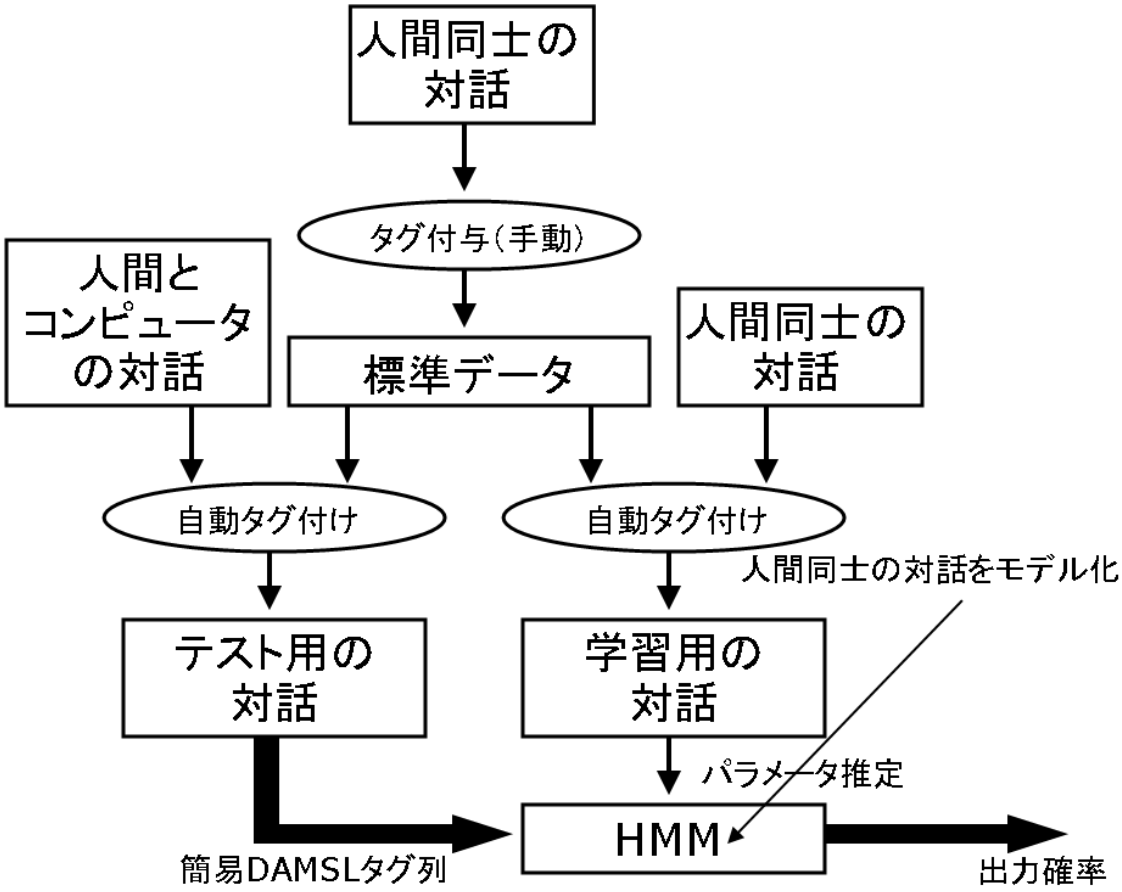

図 3.1　HMM による対話評価法における処理の流れ

対話は人間同士の対話 118 対話 (**表 3.1** の (1))，人間と対話エージェントとの対話 48 対話 (**表 3.1** の (2)(3)) より構成されており，前章と同様，以下の制約を課した．

- 30 分のテキスト対話とする
- 1 対 1 で交互に発話する
- 話題は制限しない
- 顔文字や方言は使用しない

対話エージェントには，ELIZA 型 KELDIC とインタビュア型 KELDIC の 2 種類を用いた．KELDIC の詳細については，それぞれ，前章の 2.3.1 節，2.3.2 節で述べた．

以後の実験では，学習用対話に 35 対話 (**表 3.1** の (1)a)，テスト用対話に 24 対話 (**表 3.1** の (1)b) を用い，手動でタグ付与された 59 対話 (**表 3.1** の (1)c) を自動タグ付与のための標準データとして使用した．さらに，全ての対話は，対話を行った被験者によって対話の自然さを 1(極めて不自然) から 5(極めて自然) の 5 段階で主観評価されており，平均

表 3.1　実験に用いた対話

対話の種類	(1) 人間同士の対話 {48}			(2)ELIZA 型 KELDIC と人間との対話 {24}	(3) インタビュア型 KELDIC と人間との対話 {24}
構成と使用目的	a. 学習用	b. テスト用	c. 標準データ	テスト用	テスト用
対話数	35	24	59	24	24
1 対話あたりの発話数の平均 (総数)	66.91 (2342)	66.83 (1604)	66.88 (3946)	87.00 (2088)	52.67 (1264)
1 対話あたりの発話数の標準偏差	15.42	12.41	14.16	52.47	23.25
1 発話あたりの形態素数の平均 (総数)	10.44 (24448)	11.25 (18050)	10.77 (42498)	8.91 (18610)	8.58 (10844)
1 発話あたりの形態素数の標準偏差	6.56	7.31	6.89	5.86	6.00
タグの種類数	19	19	49	23	16
対話の自然さ	4.12			1.67	2.29

{} 内は被験者数を表す

値が**表 3.1** に示されている．アンケートにおける対話の自然さとは，被験者自身の応答も含めた対話全体の自然さを被験者が主観評価したものである．この結果から，インタビュア型 KELDIC は ELIZA 型 KELDIC より自然な対話が可能であるといえる．しかしながら，インタビュア型 KELDIC の自然さは人間と比べて 1.83 低く，人間らしい知的な応答を行うレベルには達していない．これは，インタビュア型 KELDIC が発話の意味を利用していないことや，発話解析の誤りによって返答の選択に誤りが生じることなどが原因として考えられる．また，当然ながら人間同士の対話は最も自然な対話と評価されている．

表 3.2　対話に対する手動タグ付与の例

発話者	発話 [タグ]
A (人間)	駅からちょっと行ったところに, 有名なたこ焼き屋さんがありますよ. [Statement]
B (人間)	たこ焼き屋さんですか！ これまた魅力的ですね！ [Understanding,Other-forward-function]
A	駅からの南北はちょっとわかりませんが, 大通り沿いにある大きな服屋さんの裏にあります. [Statement]
B	インターネットで探してみます！ お店の名前ご存じですか？ [Committing-speaker-future-action,Question]

表 3.1 の対話のうち，標準データ以外の対話に対しては自動でタグを付与した．自動タグ付与は標準データを元にしているため，標準データの対話に比べてその他の対話はタグの種類数が少なくなった．上で述べたように，一つの発話に対して複数のタグを付与する場合もあるため，タグの組み合わせを考慮にいれると全ての対話を通してのタグの総数は 49 種類であった．

人間同士の対話で出現頻度が上位 10 位までのタグについて，各対話における相対頻度を**図 3.2** に示す．

図より，どの対話も [Statement][Question] が大部分を占めていることがわかる．各対話の上位 3 種類のタグを比較すると，どの対話も上位 3 種類のタグで全体の 80% 以上を占めているが，インタビュア型 KELDIC と人間との対話では，[Understanding] が 3 位のタグになっている点で違いがみられる．これは，インタビュア型 KELDIC が相手の意見に理解を示す「なるほど」といった相槌を頻繁に使用するためと考えられる．タグを手動で付与した前章の場合と比較すると，タグの出現頻度の傾向は類似しているが，対話エージェントにおいて [Statement,Question] が減少している．これは，標準データにおいて [Statement,Question] の出現頻度が少なかったためと考えられる．

なお, 全対話を通して最も多いタグの組み合わせは, [Statement, Question]（361 回）である．他には，[Statement, Uninterpretable]（100 回）や，[Statement, Understanding]

（57 回）などがタグの組み合わせとして頻出する．

HMM を用いた類似度の計算

本研究では，HMM によって人間同士の対話をモデル化し，その HMM に評価したい対話を入力することにより対話の自然さを評価する．

本手法では発話毎に付与された簡易 DAMSL タグを出力シンボルとして用いる．学習系列には人間同士の対話を簡易 DAMSL タグ系列で記述したものを用い，人間同士の対話を反映したモデルを作成する．学習系列が複数あるため，学習系列の出力確率の総積が局所的最大になるようにモデルパラメータを求める [53]．学習系列には人間同士の対話を用いているので，出力確率は入力系列，すなわち評価したい対話と人間同士の対話の類似度と考えることができる．

3.3　タグ付与の自動化

人手による処理では，意味的に正確なタグを付与することができるが，大量の発話を処理することは難しいという問題がある．一方，コンピュータによる自動付与の精度は，人手による付与ほどの精度はないが，大量の発話を処理することができる [55]．

対話の評価を HMM を用いて行うためには，評価したい対話に対してタグを付与する必要がある．しかし，人手によって全ての対話に対してタグを付与することは非現実的であり，コンピュータによるタグの自動付与が必要となる．

本研究では，以下の 5 種類の自動タグ付与手法を比較した．

- DICE 係数を用いた手法
- 情報量を用いた手法
- Naive Bayes を用いた手法
- SVM を用いた手法
- CRF を用いた手法

3.3.1　DICE 係数を用いた自動タグ付与手法

本手法では，タグを付与したい発話と，標準データ内の全ての発話との類似度を計算し，最も類似度が高い発話と同じタグを付与する（**図 3.3**）．DICE 係数を用いた手法と，情報量を用いた手法は，発話に含まれる単語の 2-gram を使用している．単語の 2-gram とは隣り合った単語の組のことである．ただし，助詞はすぐ前の単語と密接に関係していると考え，助詞とその直前の単語をまとめて一つの単語として扱う．

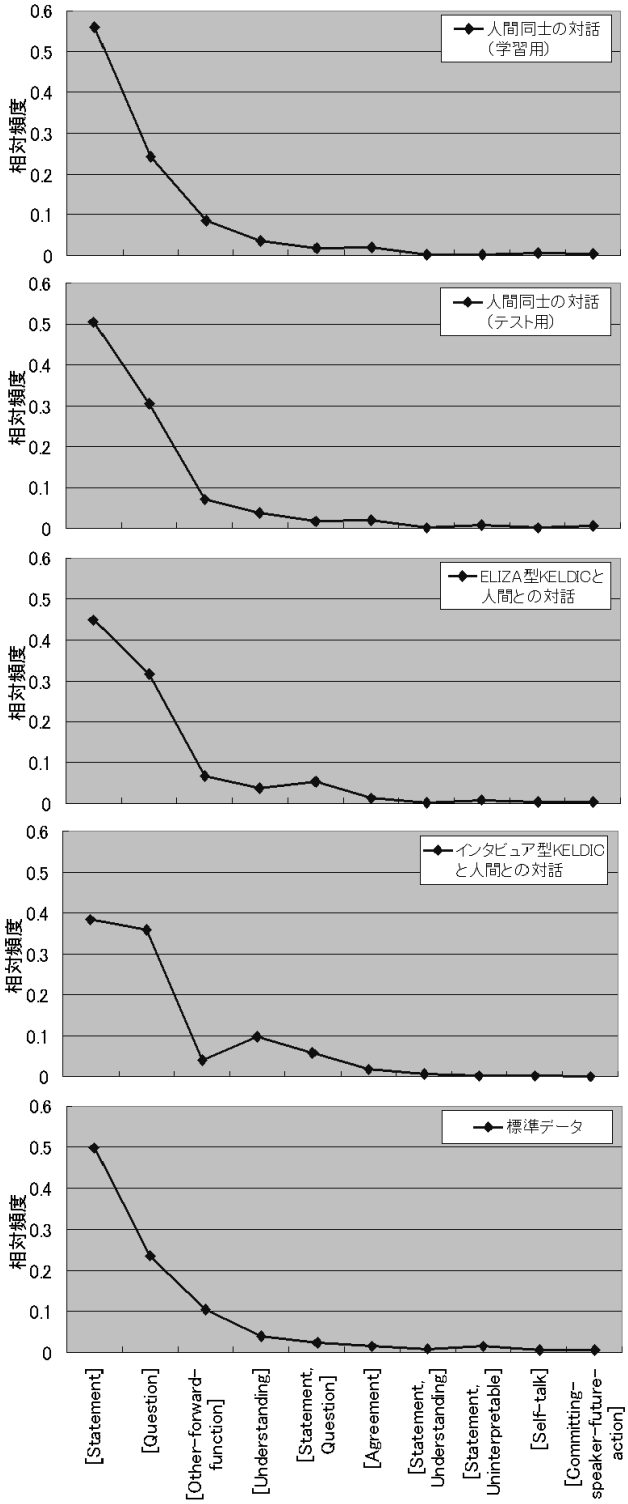

図 3.2　対話毎のタグの出現頻度

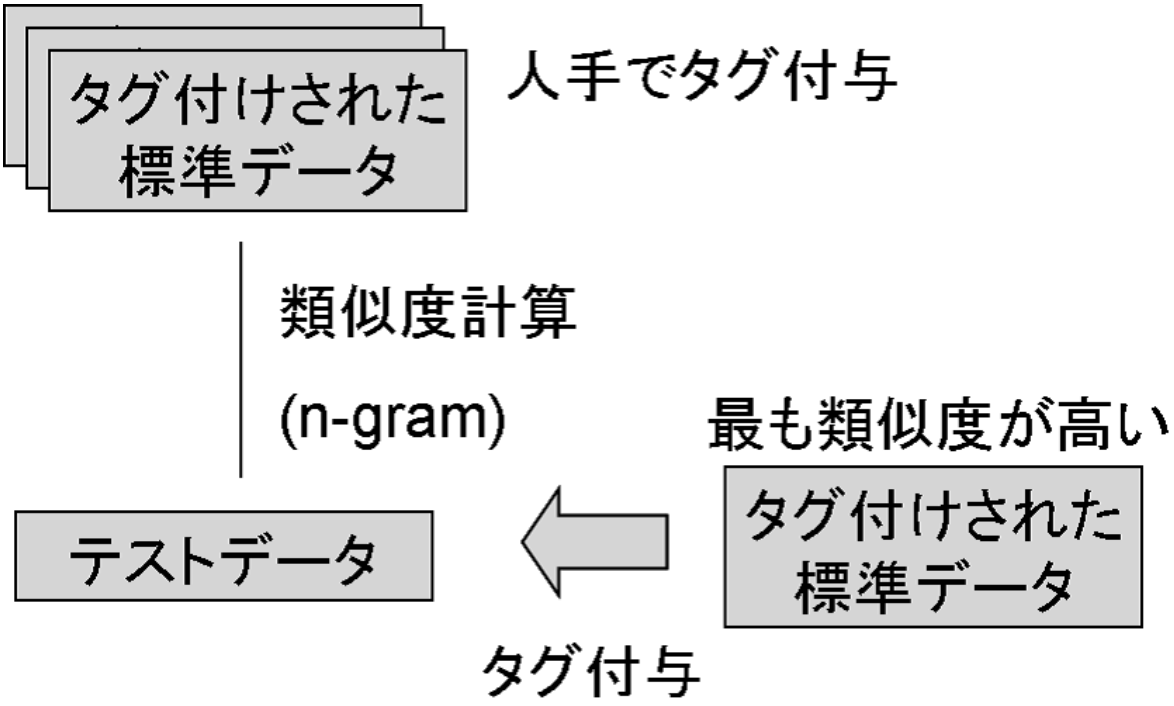

図 3.3　DICE 係数，情報量による自動タグ付与の概要

たとえば，「それで良いですよね？」「大学ですよね？」という 2 つの発話であれば，2-gram はそれぞれ

- 「それ-で」，「で-良い」，「良い-ですよね」，「ですよね-？」
- 「大学-ですよね」，「ですよね-？」

となる．発話中の単語を抽出するため，ChaSen [56] を用いて発話を形態素解析する．

2 つの発話に含まれる単語の 2-gram の集合をそれぞれ A,B で表すとき，類似度は以下の計算によって求められる．

類似度の計算には，発話に含まれる単語の 2-gram の数を用いる．

$$\text{類似度} = \frac{2 \times |A \bigcap B|}{|A| + |B|} \tag{3.1}$$

$|*|$ は $*$ に含まれる 2-gram の数を表す．この計算で求められる類似度のことを DICE 係数という [57].

たとえば，「それで良いですよね？」「大学ですよね？」という 2 つの発話の類似度を計算する．2-gram はそれぞれ

- 「それ-で」，「で-良い」，「良い-ですよね」，「ですよね-？」4 個
- 「大学-ですよね」，「ですよね-？」2 個

であり，「ですよね-？」１個が一致している．よって，

$$\begin{aligned} \text{類似度} &= \frac{2 \times 1}{4 + 2} \\ &= \frac{1}{3} \end{aligned}$$

となる．

3.3.2 情報量基準による自動タグ付与手法

発話があるタグ c_k をもつ尤度 a_k を，以下のように定める．

$$a_k = w_k + \sum_{m=1}^{M} O_m w_{mk} + \sum_{m=1}^{M} \sum_{n=1}^{M} O_{mn} w_{mnk} \tag{3.2}$$

ただし，

M：単語の種類の数
O_m：単語 v_m が発話内に含まれているとき 1, それ以外 0
O_{mn}：単語 v_m と単語 v_n の並びが発話内に含まれているとき 1, それ以外 0

$$w_k = \log P(c_k) \tag{3.3}$$
$$w_{mk} = I(v_m, c_k) \tag{3.4}$$
$$w_{mnk} = I(v_m v_n, c_k) - I(v_m, c_k) - I(v_n, c_k) \tag{3.5}$$
$$I(A, B) = \log \frac{P(A|B)}{P(A)} \tag{3.6}$$

である．ここで $P(x)$ は x の出現確率を表す．例えば，$P(c_k)$ はタグ c_k が標準データに現れる確率を表す．v_m は「私」,「今日」のような単語を表し，$v_m v_n$ は 2-gram を表す．

全てのタグについて尤度を計算し，尤度が最大となるタグを付与する．

3.3.3 Naive Bayes による自動タグ付与手法

自動タグ付与の手法として，Naive Bayes が，Grau らによって提案されている [58]. この手法では，以下の式に基づき，発話 x にラベル y を付与する．

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, P(y|x) \tag{3.7}$$
$$= \underset{y}{\operatorname{argmax}}\, \frac{P(y)P(x|y)}{P(x)} \tag{3.8}$$

ここで，$P(x)$ は y に関係ないため，

$$\hat{y} = \underset{y}{\operatorname{argmax}} P(y)P(x|y) \tag{3.9}$$

$$= \underset{y}{\operatorname{argmax}} P(y) \prod_{i=1}^{I} P(v_i, v_{i-1}, \ldots, v_{i-n+1}|y) \tag{3.10}$$

I は発話中の n-gram の総数である．v_i は発話内の i 番目の単語を表し，v_i, v_{i-1} は 2-gram を表す．データ量に偏りがある場合に対して，上式において，ラベルの出現確率 $P(y)$ が一様であると仮定した以下に示す学習器も提案されている．

$$\hat{y} = \underset{y}{\operatorname{argmax}} P(y|x) \tag{3.11}$$

$$= \underset{y}{\operatorname{argmax}} \prod_{i=1}^{I} P(v_i, v_{i-1}, \ldots, v_{i-n+1}|y) \tag{3.12}$$

本研究では式 (3.12) を用いてラベルの付与を行っている．2-gram を使用するため，$n = 2$ とする．

3.3.4　SVM による自動タグ付与手法

クラス分類，回帰，新規性検出などの分野でよく使われている Support Vector Machine(SVM)[59] を用いてタグ付与を行う．タグ付与を行うため，SVM では線形モデルを用いて 2 値分類問題を解く．

$$y(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b \tag{3.13}$$

$y(\mathbf{x}) > 0$ と $y(\mathbf{x}) < 0$ で $\mathbf{x}$ を 2 値分類する．2 値分類された学習データ X に対して，正しい分類を行う関数 y は複数考えられるが，マージンという概念を用いて，汎化誤差が最も小さくなるような y を求める．マージンとは，図 3.4 に示すように，分類境界と学習データの間の最短距離のことである．言い換えると，マージンは分類境界と最も近くのデータ点 (サポートベクトル) までの距離として定義される．マージンを最大化するという基準を設けることで，分類境界を一意に定めることができる．

ここで，$\phi(\mathbf{x})$ はある固定された特徴空間変換関数であり，b はバイアスパラメータである．SVM では，ϕ をサポートベクトルに対するカーネル関数 $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \phi(\mathbf{x}')$ で置き換え，

$$y(\mathbf{x}) = \sum_{n=1}^{N} w_n k(\mathbf{x}, \mathbf{x_n}) + b \tag{3.14}$$

とする．カーネル関数を用いることにより，データ間の内積を定義できれば，特徴量を陽に与えることなくデータを分類することが可能となる．

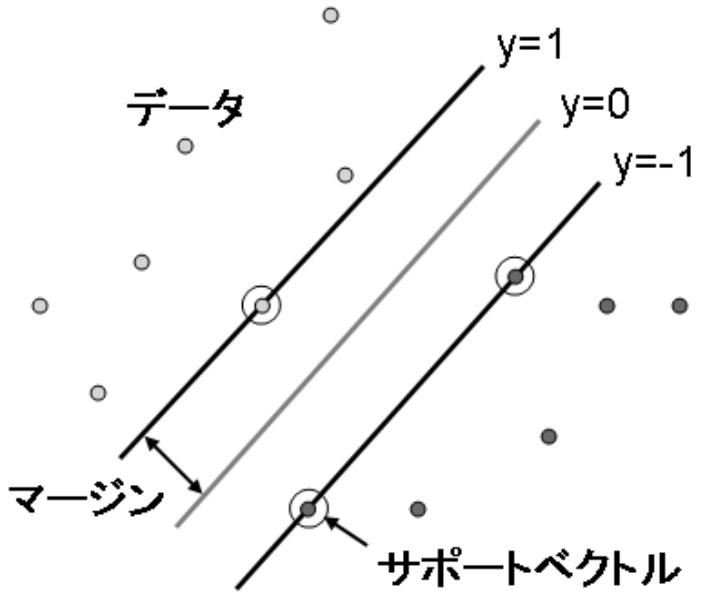

図 3.4　SVM におけるマージン

一般的に，SVM は 2 値分類問題を対象としており，自動タグ付与のような多クラス分類問題に対して適用する際には拡張が必要となる．本研究では，1 対 1 方式と呼ばれるアプローチで多クラス分類を行う．この手法は多数決方式である．すなわち，全てのクラスの組み合わせについて 2 クラス SVM を学習し，その結果得られた分類器を適用して最も多くの分類器が正例として投票したクラスを分類結果として用いる [60].

SVM の特徴量は 1-gram と 2-gram を用いた．後述する CRF とは異なり，SVM ではラベルを特徴量として用いることが困難なため，一つ前のタグは用いていない.

3.3.5　CRF による自動タグ付与手法

Conditional Random Fields(CRF) は系列ラベリング問題を解くのに適した識別モデルであり，入力発話系列 $\mathbf{x}$ と簡易 DAMSL タグ系列 $\mathbf{y}$ の対応関係を条件付確率 $P(\mathbf{y}|\mathbf{x})$ で表現する [61]．系列データに関するラベル付与問題では，HMM[62][63] が用いられ，一定の成功を収めてきた．しかし，HMM は生成モデルであり，入力系列とラベル系列の同時確率 $P(\mathbf{x}, \mathbf{y})$ を最大化することでラベルの自動付与を行う．すなわち，HMM では，以下の式に従ってラベル系列 $\hat{\mathbf{y}}$ を求める.

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} P(\mathbf{y}|\mathbf{x}) \tag{3.15}$$

$$= \underset{\mathbf{y}}{\operatorname{argmax}} \frac{P(\mathbf{x}, \mathbf{y})}{P(\mathbf{x})} \tag{3.16}$$

しかし，実際には $P(\mathbf{x})$ は $\mathbf{y}$ に関係ないため，

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\, P(\mathbf{x}, \mathbf{y}) \tag{3.17}$$

となる．このように，HMM は $P(\mathbf{x}, \mathbf{y})$ を最大化することによってラベル系列を最尤推定している．一方，CRF では，直接 $P(\mathbf{y}|\mathbf{x})$ を最大化することでラベルの自動付与を行うため，一般的に性能が良いといわれている．

自動タグ付与では，$\mathbf{x}$ と $\mathbf{y}$ は同じ長さであり，例えば長さ n の入力発話系列では，$\mathbf{x} = x_1 x_2 \cdots x_n$，$\mathbf{y} = y_1 y_2 \cdots y_n$ である．x_i は 1 つの発話を表し，y_i はその発話に付与する簡易 DAMSL タグを示す．特徴量を表現するため，CRF では素性関数 $\mathbf{f}$ を用いる．また，$\mathbf{f}$ に対応する重みパラメータを λ で表し，k 番目の特徴量に対する素性関数を f_k，重みパラメータを λ_k とする．このとき，$P(\mathbf{y}|\mathbf{x})$ は，次式で表される．

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_x} exp\Big(\sum_{i=1}^{n} \sum_{k} \lambda_k f_k(\mathbf{x}, \mathbf{y}, i) \Big) \tag{3.18}$$

ただし，Z_x は全系列の総和を 1 にするための正規化項であり，

$$Z_x = \sum_{\mathbf{y}} exp\Big(\sum_{i=1}^{n} \sum_{k} \lambda_k f_k(\mathbf{x}, \mathbf{y}, i) \Big) \tag{3.19}$$

となる．パラメータ λ は，最尤推定で求めることができる [64]．素性関数は，0, 1 の 2 値を返す関数が用いられる．

他の手法とは異なり，CRF を用いた手法では，複雑な特徴量を扱うことが可能である．本手法では，以下に示す 3 種類の特徴量を，タグ毎に異なる素性関数で表現した．

1. 標準データに 2 回以上出現した 1-gram
2. 標準データに 2 回以上出現した 2-gram
3. 一つ前のタグ

すなわち，1,2 では x_i の中に，ある 1-gram（または 2-gram）が存在するとき 1 を返す素性関数を定義する．また，3 では $y_{i-1} y_i$ が特徴量としたタグの並びと一致するとき 1 を返す素性関数を定義する．

図 3.5 で，「それで良い」という発話を例にとって特徴ベクトルを説明する．まず，1-gram として単語全てを抽出する．この場合，「それ」「で」「良い」が 1-gram に相当する．全ての単語が標準データに 2 回以上出現しているため，「それ」「で」「良い」を特徴量とする．次に，2-gram として連続する 2 単語を抽出すると，「それ-で」「で-良い」が 2-gram に相当する．この中で，「それ-で」は標準データに 2 回以上出現しているため，特徴量とする．一方，「で-良い」は 1 回のみの出現であるため，特徴量から除く．また，一

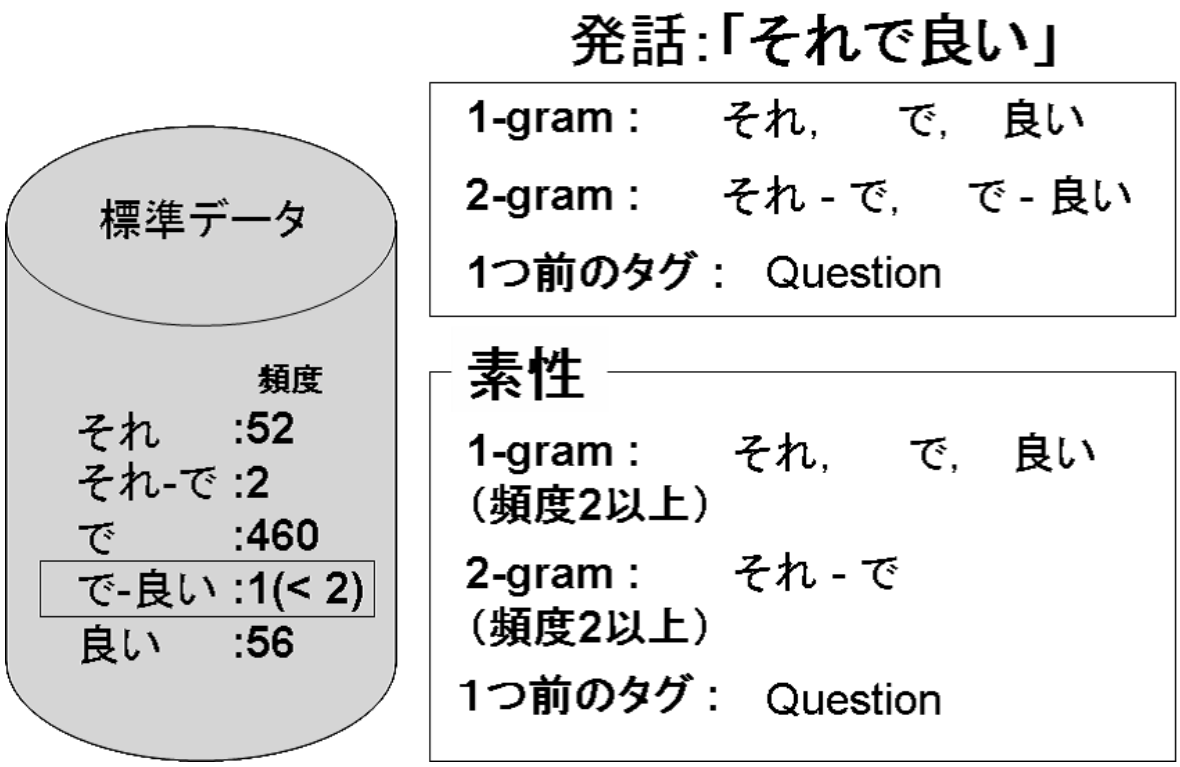

図 3.5　特徴量の例

つ前の発話に「Question」が付与されたとすると，「Question」も特徴量とする．もし，「それで-良い」に「Statement」タグが付与されるとすると，「Statement」の素性関数の中で，上記の特徴量を表す素性関数の値が 1 となる．ただし，付与されるタグは明らかではないため，全てのタグの組み合わせを考慮する必要がある．

以上を用いて，入力発話系列 $\mathbf{x}$ に対して，最適な簡易 DAMSL タグ系列 $\hat{\mathbf{y}}$ は，

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\, P(\mathbf{y}|\mathbf{x}) \tag{3.20}$$

により求めることが出来る．ここで，$\hat{\mathbf{y}}$ の計算には Viterbi アルゴリズムを用いる．

3.4　自動タグ付与の評価実験

3.4.1　実験の概要

以下では，**3.3.5** で紹介した自動タグ付与の評価実験について述べる．実験に使用した対話は標準データ (**表 3.1(1)c**) である．

自動タグ付与の性能評価には 59 分割交差検証法を用いた．すなわち，手動でタグを付与した 59 対話のうち，58 対話をパラメータ推定のための学習データ，残り 1 対話をタグ

表 3.3　自動タグ付与の結果

手法	完全一致率 (%)	部分一致率 (%)	κ
DICE 係数	56.03	63.13	0.5565
情報量	46.93	83.15	0.4221
Naive Bayes	60.14	72.12	0.4379
SVM	66.95	73.75	0.4943
CRF	75.77	83.20	0.6371

未知の対話とみなして自動タグ付与を行い，手動で付与したタグと比較した．タグ未知とみなす対話を変更し，59 通り実験を行い，正しくタグ付けできた発話数の平均によって，自動タグ付与の性能を評価した．

3.4.2　実験結果

実験結果を**表 3.3** に示す．

完全一致率は，自動付与したタグと手動付与したタグとが完全に一致した割合である．部分一致率は，複数のタグのうち，一部が一致しているような場合，たとえば正解が「Question + Statement」である発話に「Question」を自動付与した場合も正解とみなしたときの割合である．また κ は κ 統計量を表し，

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{3.21}$$

である．$P(A)$ はタグの完全一致を表す．$P(E)$ はタグが偶然一致する確率，すなわち

$$P(E) = \sum_k P_m(v_k) * P_a(v_k) \tag{3.22}$$

である．ここで，v_k は各タグ，$P_m(v_k)$ は標準データにおけるタグ v_k の出現確率，$P_a(v_k)$ は自動タグ付与したデータにおけるタグ v_k の出現確率を表す．この結果より，CRF は，他の手法よりも精度が高いことがわかった．これは，CRF の識別精度の高さだけでなく，特に CRF が一つ前のタグを特徴量として用いているためと考えられる．CRF は，素性関数 $f_k(\mathbf{x}, \mathbf{y}, i)$ から明らかなように，特徴量としてラベル $\mathbf{y}$ を用いることが容易である．一方，DICE 係数，情報量，Naive Bayes，SVM ではラベルを特徴量として実装することが困難であるため，一つ前のタグを特徴量として用いていない．これにより，CRF は対話の流れによるタグの変化に対応できたのではないかと考えられる．

表 3.2 と同じ対話に対して，自動タグ付与した結果を**表 3.4** に示す．この例の場合，**表**

表 3.4　対話に対する自動タグ付与の例

発話者	発話 [タグ]
A (人間)	駅からちょっと行ったところに, 有名なたこ焼き屋さんがありますよ. [Statement]
B (人間)	たこ焼き屋さんですか！ これまた魅力的ですね！ [Understanding,Statement]
A	駅からの南北はちょっとわかりませんが, 大通り沿いにある大きな服屋さんの裏にあります. [Statement]
B	インターネットで探してみます！ お店の名前ご存じですか？ [Committing-speaker-future-action,Question]

3.2 の手動による付与の場合と比較すると，「たこ焼き屋さんですか！これまた魅力的ですね！」という発話に対して手動タグ付与では [Understanding,Other-forward-function] が付与されているのに対し，自動タグ付与では [Understanding,Statement] と異なったタグが付与されている．それ以外の 3 つの発話には同じタグが付与されている．異なったタグが付与された原因は，大きく分けて以下の 2 つである．

- 標準データ内に一致する単語が無かった場合
- 標準データ内に同じ発話でも異なるタグが付与されている場合

本手法では発話に含まれる単語を特徴として計算しているので，標準データに存在しない単語が現れた場合は最適なタグを選択することができず，タグの付与は失敗する．これらの問題を解決するためには，自動タグ付与に用いる標準データ数を増やす必要がある．

また，CRF は系列全体の最適化を行うため，対話の進行中にリアルタイムでタグ付与を行う場合，過去の発話に対して付与されたタグが他のタグに変化する可能性がある．今回の実験では収録済みの対話に自動タグ付与を行うため問題はないが，リアルタイムにタグを付与する場合は注意が必要である．

次節では，本節の実験で最も高精度であった CRF を用いて自動タグ付与を行う．

3.5　自動タグ付与を用いた対話評価法の評価実験

3.5.1　HMM による対話の評価手法

自動タグ付与の対話評価法に対する影響を調査するため，前章の手動タグ付与による評価法を基準とし，学習用対話に対して自動タグ付与をした場合の評価と比較する．人間同士の対話の学習により HMM を作成し，人間と対話エージェントとの対話を入力したときの出力確率を調べる．また，参考のため人間同士の対話についても評価する．HMM 出力シンボルには簡易 DAMSL タグ，学習用対話には人間同士の対話のうち 35 対話 (**表 3.1**(1)a) を用いる．なお，全ての学習用対話には，前節で最も高精度であった CRF による自動タグ付与手法を用いて簡易 DAMSL タグを付与した．HMM の状態数を 1 から 20 まで変化させ，最もエントロピーが低くなるモデルを採用する．HMM のパラメータを λ とすると，エントロピー $H(\lambda)$ は以下の式で求められる．

$$H(Q_i) = -\sum_{k}^{K} p(v_k|Q_i) log_2 p(v_k|Q_i) \tag{3.23}$$

$$H(\lambda) = \sum_{i}^{N} \omega_i H(Q_i) \tag{3.24}$$

ただし，N，K はそれぞれ状態数とシンボル数を，$p(v_k|Q_i)$ は状態 Q_i でシンボル v_k を出力する確率を表す．ここで，ω_i は定常状態確率を表し，十分な遷移の後で状態 i に存在する確率を表す．エントロピーが小さいほど，HMM が学習用対話の特徴をよく表していると考えられる．

HMM の状態数とエントロピーの関係を図 3.6 に示す．エントロピーが最小となった状態数は 11 であった．

前章と同様に，学習後の HMM の内部構造を**図 3.7** に示す．前章との違いは学習データに自動タグ付与を行っている点である．**図 3.7** より，初期状態 Q_0 では挨拶を表す「Other-forward-function」を高確率で出力し，また質問を表す「Question」を出力しやすい状態 Q_5 では，思ったことや説明などを表す「Statement」を高確率で出力する状態 Q_3 に遷移することがわかる．これは，質問-返答といった対話の流れを示していると考えられる．さらに，「Statement」を高確率で出力する状態は多くの状態から高確率で遷移があり，対話における中心を担っていると考えられる．

このことから，手動タグ付与を用いた場合と同様に，自動タグ付与を用いた場合でも，人間同士の対話を学習した HMM は，対話の始めに行われる挨拶や，質問に対する返答といった人間同士の対話で見られる構造をモデル化できているといえる．

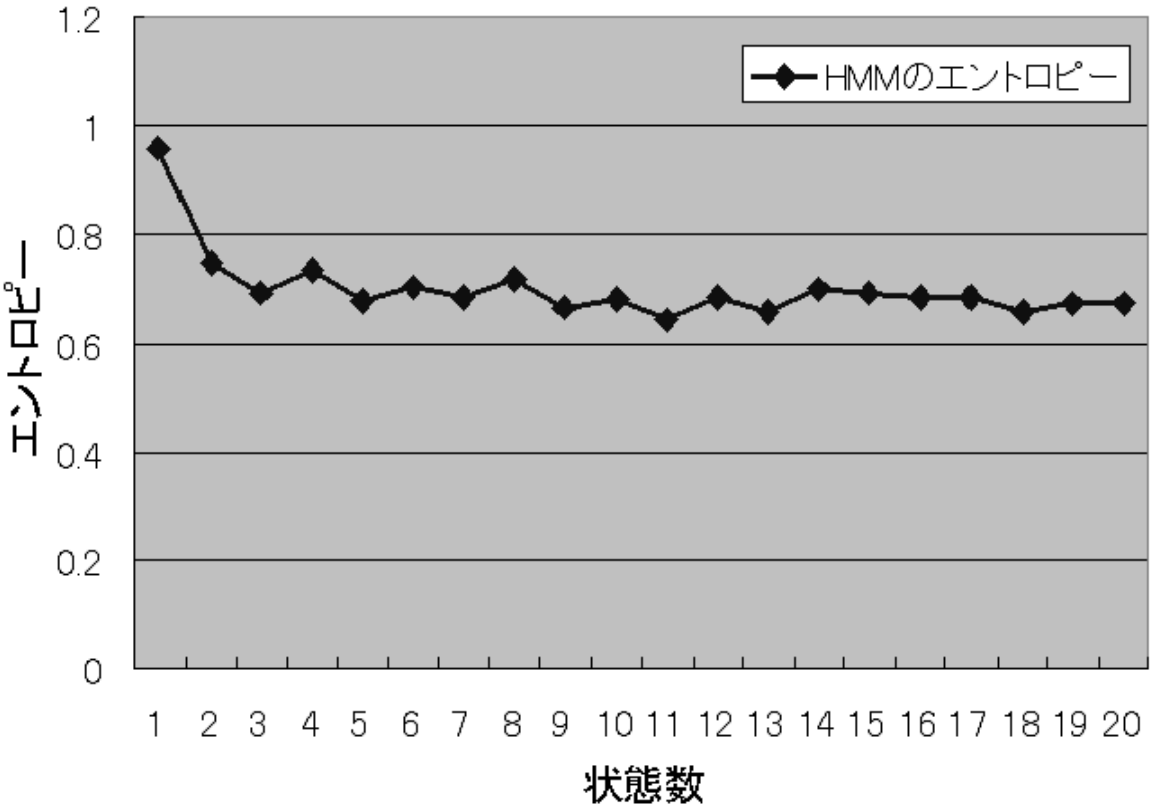

図 3.6　HMM の状態数とエントロピーの関係

まず，HMM の出力確率と人間の主観評価を比較することにより，人間の主観と同様に自然な対話，不自然な対話の評価が可能かどうかを調べる．

次に，**3.2.2** で述べた ELIZA を模して作られた ELIZA 型 KELDIC と，プロのインタビュアを模したインタビュア型 KELDIC を比較評価した．使用する対話は**表 3.1** の通りである．全ての対話は **3.3.5** で述べた自動タグ付与手法によりタグ付与した．

また，自動タグ付与が HMM を用いた評価法に与える影響を確認するため，テストデータに手動タグ付与を行った場合と比較した．同様に，学習データとテストデータの両方に手動タグ付与を行った前章の実験結果と比較した．

3.5.2　実験結果

図 3.8 は (1)ELIZA 型 KELDIC と人間との対話 (**表 3.1**(2))，(2) インタビュア型 KELDIC と人間との対話 (**表 3.1**(3))，(3) 人間同士の対話 (**表 3.1**(1)b) をそれぞれ HMM に入力したときの評価結果である．横軸は HMM による出力確率を表し，縦軸はその相対頻度を表す．同様に，**図 3.9** は比較のため同じ対話を使用し，学習データに自動タグ付与，テストデータに手動タグ付与を用いた場合の HMM による対話の評価結果で

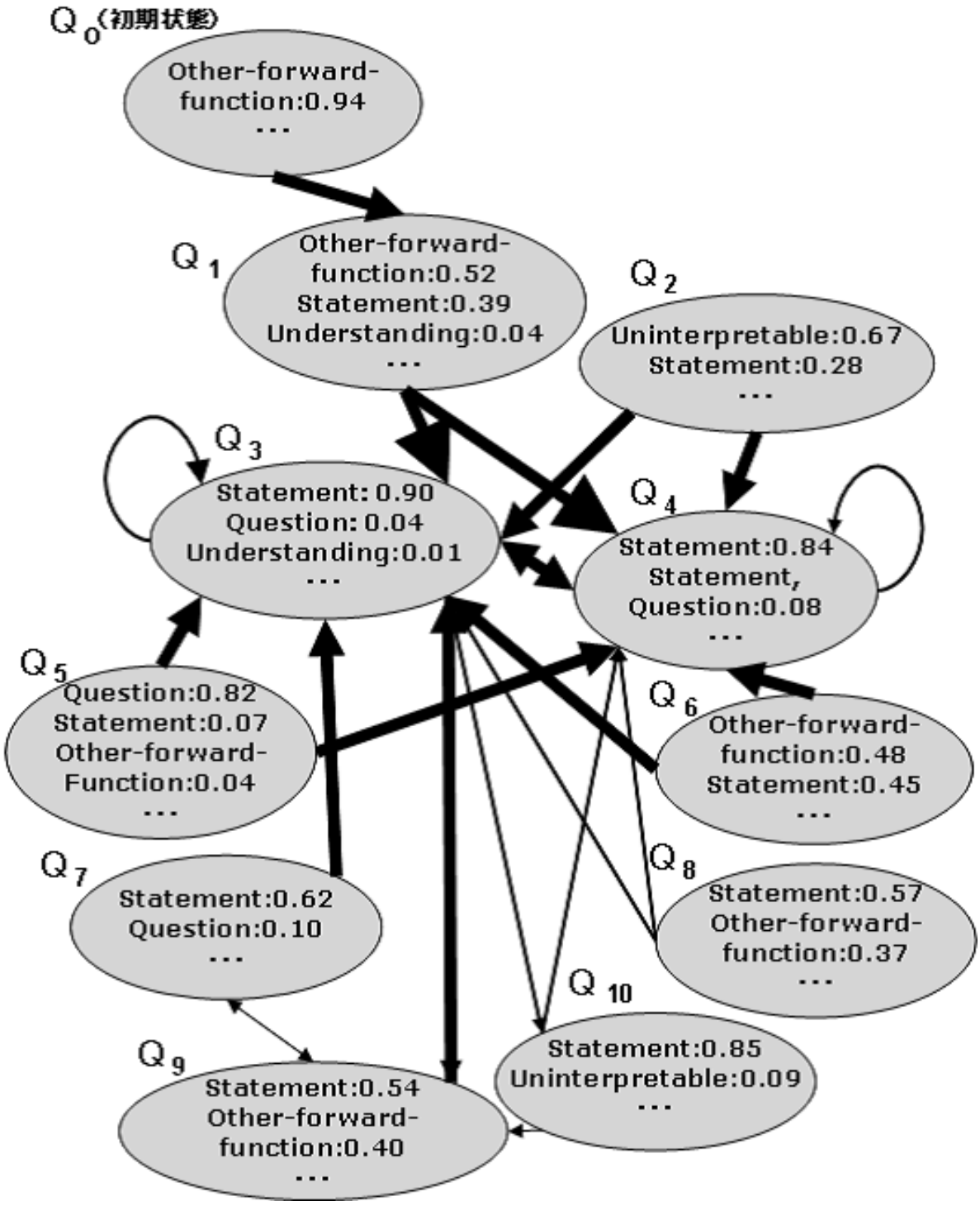

図 3.7　自動タグ付与を用いた HMM の構造

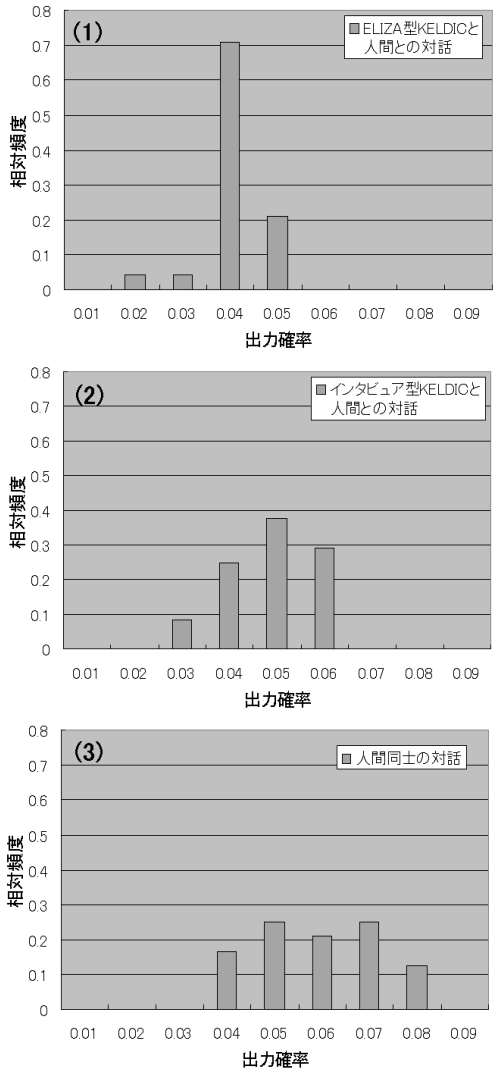

図 3.8　自動タグ付与 +HMM による対話エージェントの性能比較

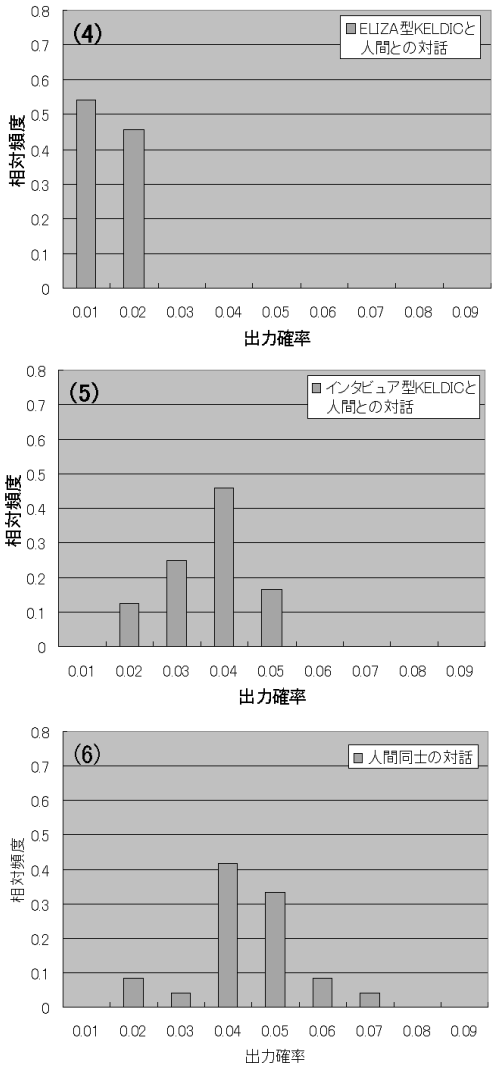

図 3.9　手動タグ付与 +HMM による対話エージェントの性能比較

ある．学習データ，テストデータの両方に対して手動タグ付与を行った場合の評価結果は前章の**図 2.6** に示されている．

まず，対話エージェントの比較評価の結果を示す．

図 3.8 から明らかなように，HMM による評価値は，(1)(2)(3) の順に高くなっており，**表 3.1** の対話の自然さと同様の傾向を示していることから，HMM による評価法が妥当であるといえる．また，**図 3.8**，**図 3.9** から，テストデータに対して手動タグ付与を行った場合と自動タグ付与で同様の傾向を示すことがわかった．これは，全ての処理を手動タグ付与で行った前章の結果と同様であった．

視覚的に比較するより定量的な評価結果を得るため，前章と同様に評価値頻度分布間の

表 3.5　クラス内分散・クラス間分散比

記号	比較する対話の種類 (数字は**図 3.8**, **3.9** のグラフ内の番号を表す)	J_σ
J_{12}	(1) と (2)	0.35
J_{13}	(1) と (3)	0.90
J_{23}	(2) と (3)	0.21
J_{45}	(4) と (5)	2.88
J_{46}	(4) と (6)	3.00
J_{56}	(5) と (6)	0.09

距離をクラス内分散・クラス間分散比 J を用いて計算する．この値が小さいほど分布間の距離は小さい．**図 3.8**(1)，**図 3.8**(3) で表される二つの分布間の距離を J_{13} とおくと，

$$J_{13} = 0.90 \tag{3.25}$$

と求められる．同様にして

- インタビュア型 KELDIC と人間との対話の評価値頻度分布 (**図 3.8**(2)) と，
- 人間同士の対話の評価値頻度分布 (**図 3.8**(3)) との距離 J_{23} は

$$J_{23} = 0.21 \tag{3.26}$$
$$< J_{13} \tag{3.27}$$

となり，インタビュア型 KELDIC の方が，より自然な対話を実現していることが定量的にも確かめられる．**図 3.9** についても同様に分布間の距離を計算できる．**表 3.5** は，J_{12} や，**図 3.9** の (4)，(5)，(6) から求めた J_{45}，J_{46}，J_{56} を，上記結果と合わせてまとめたものである．

表 3.5 の中で，J_{13}，J_{23}，J_{46}，J_{56} は理想的な対話 (人間同士の対話) からの距離を示し，値が小さい程良い．一方 J_{12}，J_{45} は評価対象となる対話エージェント同士の距離であり，値が大きい程，対話エージェントの性能に差があることを示す．主観評価により対話エージェントの性能に大きな差が見られるならば，J_{12}，J_{45} の値も大きいことが望ましい．

表 3.1 の主観評価と比較すると，J_{12}，J_{13}，J_{23} は $J_{12} < J_{13}$ かつ J_{23} が 0.21 と小さい値であり，同じ傾向を示している．また，テストデータを手動タグ付与した場合である

J_{45}，J_{46}，J_{56} と，全ての処理を手動タグ付与で行った前章の場合も同じ傾向を示している．ここで，J_{12} と J_{45} では値が大きく異なっている．同様に，**図 3.8**(1)，**3.9**(4) では，結果が大きく異なるが，この主な理由として，ELIZA 型 KELDIC が人間同士の対話には現れにくい「続けてください」といった特殊な発話をすることが挙げられる．自動タグ付与は，人間同士の対話に手動で付与されたタグを学習しているため，このような人間同士の対話に現れにくい特殊な発話に対してはタグ付与の誤りがおこりやすい．しかし，正しくタグ付与された発話系列であっても理想的な対話の構造から離れているため，評価の傾向に差が現れるほどの影響はないと考えられる．

以上の結果より，自動タグ付与を用いて人間同士の対話を学習した HMM は，自然な対話をモデル化できていると考えられる．また，自動タグ付与と手動タグ付与で，評価の傾向に差がないことが確認された．人間による自然さの主観評価 (**表 3.1**) と HMM による評価 (**図 3.8**) の比較を**図 3.10** に示す．グラフ上の各点は 1 つの対話 (**表 3.1**(1)b, (2)，(3)) に対応している．ピアソンの積率相関係数は 0.44 であり，人間による主観評価と HMM による評価との間に比較的高い相関がみられた．以上の結果は全て，タグ付与を手動で行った場合と同様であり，自動タグ付与を用いた対話評価法は有効であることが確認できた．

3.6 要約

前章で提案した非タスク指向型対話エージェントの客観的・定量的評価法では，手動タグ付与を用いていたため，大量の対話データを評価することが困難であった．そこで，評価法の有用性を高めるため，自動タグ付与手法を提案した．自動タグ付与手法として，DICE 係数を用いた手法，情報量を用いた手法，Naive Bayes を用いた手法，SVM を用いた手法，CRF を用いた手法の 5 種類を提案し，比較実験を行った．実験の結果，CRF を用いた手法が最も精度が高く，正解率は 75.77%，意味的に近いものを許容した場合で 83.20% であった．また，κ 統計量は 0.6371 であった．以上のことから，CRF を用いた自動タグ付与の有効性が確認できた．

そこで，CRF を自動タグ付与手法として用い，HMM による対話の評価を行った．まず，手動でタグを付与した標準データを用いて，CRF のパラメータを学習した．CRF を用いて各発話に自動タグ付与を施し，全ての対話をタグ系列として表現した．そして，人間同士の対話のタグ系列を学習した HMM を作成した．評価すべき対話をこの HMM に入力し出力確率を計算することにより，対話の自然さを評価した．実験の結果，より人間に近い自然な対話を行う対話エージェントほど HMM の出力確率が高くなることを確認した．さらに，人間同士の対話は人間と対話エージェントとの対話より HMM の出力確率が高くなり，その中でも自然な対話ほど出力確率が高くなることを確認した．これは自

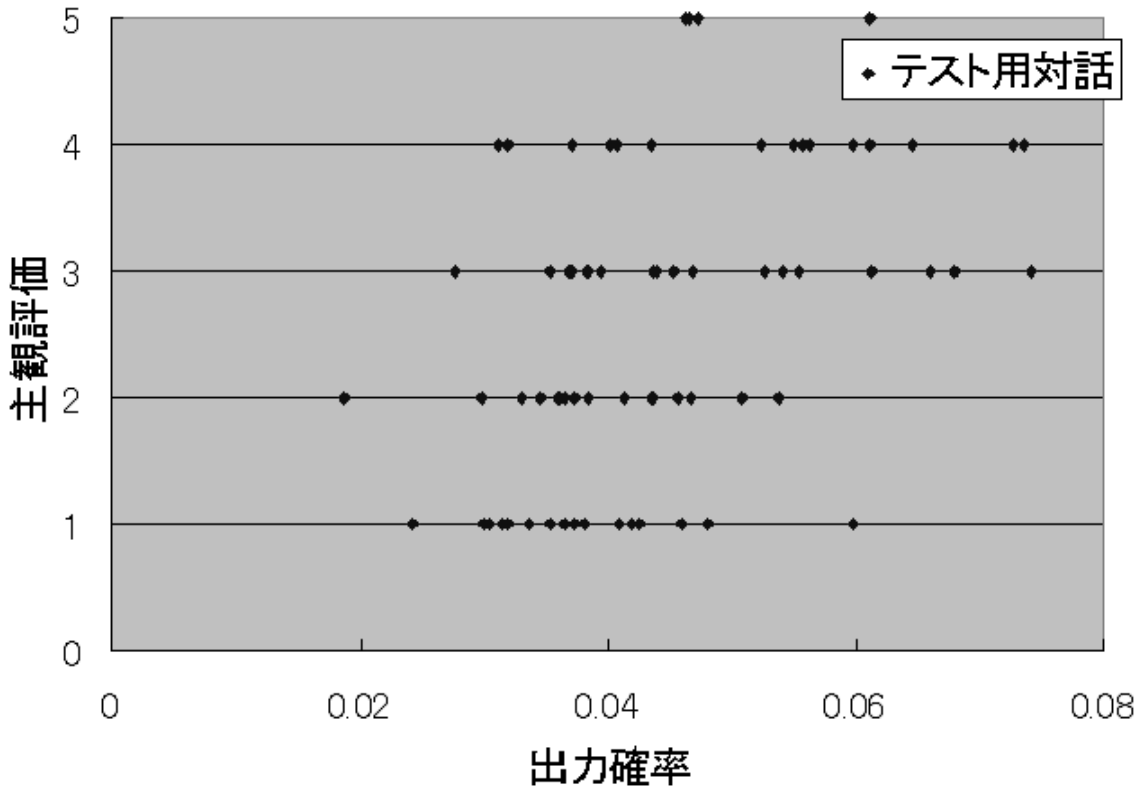

図 3.10　HMM による評価と主観評価との比較

動タグ付与を行わず，手動タグ付与のみで対話を評価した際の結果と同等の結果であり，本手法の有用性を確認した．

提案手法の性能をより向上させる方法として，簡易 DAMSL タグの自動付与手法の改良が考えられる．本研究では識別モデルである CRF を使用してタグ付与を行った．しかし，近年では，生成モデルが再評価されつつあり，特に，ノンパラメトリックな手法が研究され，急速に発展してきている [65][66]．この手法を応用することで，自動タグ付与の精度を高めることができると期待される．さらに，本評価法を用いて対話エージェントを設計・改良していくことが重要な課題である．

次章では，発話集を自動生成し，その中から適切な応答を選んで対話を行う，非タスク指向型対話エージェントを実現するための応答自動選択法について述べる．

第 4 章

発話の自動選択

4.1 はじめに

前章では，テキスト対話を対象とし，非タスク指向型対話エージェントを HMM により定量的に自動評価する方法を提案した．これにより，対話エージェントの自然さを比較評価することが可能となった．そこで，次の段階として，対談番組の司会者のように，対話を盛り上げる対話エージェントの設計を目指す．優秀な司会者は，対話相手に関する情報を基に質問を練り上げる等，周到に準備して対話に臨んでいる．本論文では，対話相手の情報から大量の発話集を用意して対話することにより，このような司会者をコンピュータ上に実現することを試みる．このような対話エージェントを実現するためには，

1. 大量の発話集の自動生成
2. 大量の発話集から最適な応答の自動選択

の 2 つの課題を解決しなくてはならない．本論文では，課題 2 を取り上げる．すなわち，対話相手のプロフィールに基づいた大量の発話候補が事前に用意されているものとし，その中から適切な応答を自動選択することを目指す．

応答の自動選択を実現することにより，対話エージェントだけでなく，人間の応答を補助するエージェントの実現も期待できる．すなわち，対談やパネルディスカッションにおける人間の司会者に対して，エージェントが状況や対話相手に応じて適切な応答候補を提案することが可能となる．

このような，ある入力に対して複数の候補の中から適切な候補を選択する問題に対しては，機械学習を用いた統計的な手法が数多く提案されてきた [67, 68, 69]．例えば，日本語の係り受け解析では，入力された文節に対して，複数の文節候補から，適切な係り先の文節を探す問題に機械学習手法を適用している．この手法では，候補となる文節を「係る」「係らない」の二値分類問題として扱い，決定木，最大エントロピー法といった手法で分

類している [67, 68]．また，2 値分類以外の手法として，優先度学習と呼ばれる機械学習手法を適用することで，候補間の相対的な大小関係をモデル化し，候補の順位付けを行う手法が提案されている [69]．

本章でも，発話候補の相対的な大小関係をモデル化することで発話候補を順位付け，発話を選択する手法を提案する．まず，発話の選択問題を定義する．その後，人間の発話選択を学習する手法について説明する．さらに，手動選択との比較実験を行い，共起情報を用いた本手法の有効性を確認する．以下，4.2 節では，発話選択法について詳細に述べる．4.3 節では，統計的発話選択法について述べる．4.4 節では，提案手法の有効性を確認するための評価実験について述べる．4.5 節では，発話の自動選択手法についての結果をまとめる．

4.2　適切な応答の選択

本手法では，対話の 1 時点を状態 s，それに対応する発話候補の集合を

$$A_s = \{a_1^s, a_2^s, \ldots, a_{m_s}^s\} \tag{4.1}$$

と定義する．式中の m_s は，状態 s のときの発話候補数である．s と A_s の例をそれぞれ**表 4.1** と**表 4.2** に示す．**表 4.1** は，1 番から始まり，8 番まで順番に発話が続いている対話の状態 s を表している．**表 4.2** の A_s は s の次に来る発話の候補である．すなわち，s の最後の発話である，「水族館の静かで落ち着いた雰囲気が好きです。」に対する応答の候補集合が A_s である．また，$m_s = 6$ である．

ここで，A_s は対話相手の情報を基に質問集を用いる手法や，s に含まれる発話に対してテンプレートを用いて発話を生成するなどの手法で作成される．ただし，すでに述べたように本研究では選択手法のみに着目し，A_s の生成法については扱わない．

ここで，s と A_s は以下の条件を満たすと仮定する．

- 全ての s について A_s を生成可能である
- s に対し，A_s は 1 個以上の適切な発話（正解事例）を含む

正解事例を含まない場合は，A_s の生成手法に問題があると考えられるため，本稿では扱わない．

適切な発話の選択は，A_s から正解事例集合 C_s を見出すことである．例えば，**表 4.1** の 8 番目の発話に対する応答を，**表 4.2** の中から選ぶとすると，「水には何か心を落ち着かせる力がありますよね？」，「水族館は落ち着きますよね。」の 2 つが正解事例である．つまり，この場合の正解事例集合 C_s は，$C_s = \{a_1^s, a_4^s\}$ となる．

表 4.1 対話の 1 時点 s の例

発話番号	発話者:発話
1	A：どんな魚、動物が好きですか？
2	B：イルカが好きです。
3	A：どうしてですか？
4	B：イルカはとても賢いからです。
5	A：私も大好きです。
6	B：それはうれしいです。
7	A：水族館の好きなところはどんなところですか？
8	B：水族館の静かで落ち着いた雰囲気が好きです。

表 4.2 対話の 1 時点 s に対応する発話候補の集合 A_s の例

番号	発話候補
a_1^s	水には何か心を落ち着かせる力がありますよね？
a_2^s	その動物のどんなところが好きですか？
a_3^s	あと４年近くありますもんね。何か見つかるといいですね。
a_4^s	水族館は落ち着きますよね。
a_5^s	社会に役立つと思います。勉強頑張って下さい。
a_6^s	そうでしたか。好きな映画俳優はいますか？

4.3　統計的発話選択法

本研究で用いる発話の自動選択法について説明する．対話の状態集合を S と表す．対話の 1 状態 $s(\in S)$ と $a(\in A_s)$ に特定の処理を施すことによって生成される n 次元の特徴ベクトルを

$$\Phi(s,a) = (x_1(s,a), x_2(s,a), \ldots, x_n(s,a)) \in \mathbb{R}^n \tag{4.2}$$

とする．ここで，$x_j(s,a)$ は特徴量を表す．例えば s の最後の発話と a に注目し，特定の単語または品詞，さらにこれらの組み合わせの有無などを $\{0,1\}$ の特徴量とすることができる．

さらに，特徴ベクトルの評価値を返す関数 $f\left(\Phi(s,a)\right)$ を定める．ここで，f は $x_j(s,a)$ に関する線形な関数とし，以下のように表す．

$$f\left(\Phi(s,a)\right) = \sum_{j=1}^{n} w_j x_j(s,a) \tag{4.3}$$

ここで，w_j は $x_j(s,a)$ に対する重みを表すパラメータであり，パラメータベクトル $\boldsymbol{w}$ として以下のように表す．

$$\boldsymbol{w} = (w_1, w_2, \ldots, w_n) \tag{4.4}$$

評価関数 f を用いて，対話の状態 s に対する最適な発話 $\hat{a}$ は，以下のように求めることができる．

$$\hat{a} = \underset{a \in A_s}{\operatorname{argmax}} f(\Phi(s,a)) \tag{4.5}$$

評価関数の値を降順にソートすることで，発話候補の順位付けを行うこともできる．

対話の状態 s_i と，それに対応する正解事例集合 C_{s_i} の組み合わせを教師データ $T = \{C_{s_1}, C_{s_2}, \ldots, C_{s_M}\}$ とする．ここで，M は教師データに含まれる正解事例集合の数である．また，誤り事例集合を $\bar{T} = \{\bar{C}_{s_1}, \bar{C}_{s_2}, \ldots, \bar{C}_{s_M}\}$ とする．$\bar{C}_s$ は A_s に含まれる発話候補のうち，C_s に含まれない発話候補の集合である．すなわち，$\bar{C}_s = A_s \backslash C_s$ である．

教師データを使い，何らかの基準（目的関数）に基づいて評価関数 f のパラメータベクトル $\boldsymbol{w}$ を推定する．

本研究では，以下に示すシグモイド関数を用いる手法及び，最大エントロピー法を用いる手法を比較した．

4.3.1 シグモイド関数を用いる手法

正解事例 $a_c \in C_s$ と誤り事例 $a_{\bar{c}} \in \bar{C}_s$ が与えられたとき，最適な評価関数 $\hat{f}$ は，

$$\hat{f}\left(\Phi(s, a_c)\right) - \hat{f}\left(\Phi(s, a_{\bar{c}})\right) > 0 \tag{4.6}$$

であることが望ましい．言い換えると，$\hat{f}\left(\Phi(s, a_c)\right) - \hat{f}\left(\Phi(s, a_{\bar{c}})\right)$ の符号が正であれば良い．すなわち，

$$\mathrm{sgn}\left(\hat{f}\left(\Phi(s, a_c)\right) - \hat{f}\left(\Phi(s, a_{\bar{c}})\right)\right) = 1 \tag{4.7}$$

と等価である．ただし，$\mathrm{sgn}(x)$ は x の符号を示す整数を返す関数であり，

$$\mathrm{sgn}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases} \tag{4.8}$$

である．

ある入力に対して適切な発話が，他の入力に対して不適切なことは頻繁に見られるため，入力毎の発話候補間の相対的な大小関係をモデル化するほうが発話の選択問題の性質を適切に表現している．そこで，教師データ毎の大小関係に注目し，最適なパラメータベクトル $\hat{\boldsymbol{w}}$ を以下のように定義する．

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\mathrm{argmax}}\, F_{\boldsymbol{w}} \tag{4.9}$$

$$F_{\boldsymbol{w}} = \sum_{\substack{s, a_c \in C_s, \\ a_{\bar{c}} \in \bar{C}_s}} \mathrm{sgn}\left(t(s, a_c, a_{\bar{c}})\right) \tag{4.10}$$

$$t(s, a_c, a_{\bar{c}}) = f(\Phi\left(s, a_c\right)) - f(\Phi\left(s, a_{\bar{c}}\right)) \tag{4.11}$$

$$= \sum_{j=1}^{n} w_j (x_j(s, a_c) - x_j(s, a_{\bar{c}})) \tag{4.12}$$

式 (4.10) の右辺に含まれる $\mathrm{sgn}(x)$ は微分できないため，最適解を求めるのは困難である．そこで，$\mathrm{sgn}(x)$ をシグモイド関数を用いて次式のように連続関数で近似する．

$$\mathrm{sgn}(x) \approx \frac{2}{1 + e^{-kx}} - 1, k > 0 \tag{4.13}$$

式 (4.10) に式 (4.13) を代入して，定数を加乗算し，目的関数 $F_{\boldsymbol{w}}$ を得る．

$$F_{\boldsymbol{w}} = \sum_{\substack{s, a_c \in C_s, \\ a_{\bar{c}} \in \bar{C}_s}} \left(\frac{2}{1 + e^{-kt(s, a_c, a_{\bar{c}})}} - 1 \right) \tag{4.14}$$

$$\Rightarrow \sum_{\substack{s,a_c\in C_s,\\ a_{\bar{c}}\in \bar{C}_s}} \left(\frac{1}{1+e^{-kt(s,a_c,a_{\bar{c}})}} \right) \tag{4.15}$$

最適解 $\hat{\boldsymbol{w}}$ は，式 (4.9)，(4.12)，(4.15) より，反復スケーリング法 IIS，GIS[70] や，準ニュートン法 L-BFGS などの手法で求めることができる [71]．本研究では，共役勾配法の一種である，SCG 法 [72] を用いた．

4.3.2　最大エントロピー法を用いる手法

最大エントロピー法はデータスパースネスに強いといわれているパラメータ計算手法である [68]．各特徴の出現割合は教師データと未知のデータで等しいと仮定し，最も均一な分布に近くなるように，パラメータを推定する．例えば，教師データがない状態では，n 個のパラメータ w_i は全て $1/n$ の等確率になる．

ここでは，過学習を防ぐため，事後確率最大化によりパラメータの正則化を行う．最大エントロピー法の目的関数 $L_{\boldsymbol{w}}$ は以下の式で示される．

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}}(L_{\boldsymbol{w}}) \tag{4.16}$$

$$\begin{aligned} L_{\boldsymbol{w}} = \sigma \sum_{s,a_c\in C_s} \Bigg(& \boldsymbol{w}\Phi(s,a_c) \\ & - \log\left(\sum_{a_{\bar{c}}\in \bar{C}_s} \exp(\boldsymbol{w}\Phi(s,a_{\bar{c}})) \right) \Bigg) \\ & - \frac{1}{2}||\boldsymbol{w}||^2 \end{aligned} \tag{4.17}$$

正則化項は $-\frac{1}{2}||\boldsymbol{w}||^2$ である．σ は学習データに対するモデルの複雑さを制御するパラメータである．σ の値は交差検証の結果などから選択する．最適解 $\hat{\boldsymbol{w}}$ は，シグモイド関数を用いる手法と同様の手法で求めることができる．本研究では，SCG 法を用いて $\hat{\boldsymbol{w}}$ を計算した．

4.4　発話選択の比較評価実験

4.4.1　実験の概要

以下では，発話選択の実験について述べる．

人間の発話選択を学習したときの，発話自動選択法の精度を確認するため，人手で評価した教師データと自動選択法による順位付けを比較した．

精度の比較のため，最大エントロピー法（$\sigma = 0.1$）による発話候補自動選択も行った．

発話候補の順位付けは，評価値を降順にソートすることで実現した．
特徴ベクトルとしては，

(1) 入力と発話候補の自立語の品詞及び付属語 2-gram の組み合わせの有無
(2) 入力と発話候補の文節の数
(3) 入力と発話候補で同一の自立語の種類と数
(4) 入力と発話候補で共起する自立語の種類と数

を用いた．これらの特徴は全て $\{0,1\}$ の 2 値で表現した．入力が**表 4.1**，発話候補が「水には何か心を落ち着かせる力がありますよね？」の際の特徴ベクトルを**図 4.1** に示す．最後の発話「水族館の静かで落ち着いた雰囲気が好きです。」のみに注目して特徴ベクトルを計算している．

(1) の 2-gram の組み合わせの有無では，入力内の 2-gram と，発話候補の 2-gram を組み合わせの有無を特徴量として用いる．ただし，自立語はあらかじめ品詞に変換する．**図 4.1** では，入力中の「水族館の」の自立語「水族館」を品詞 (名詞) で置き換えて，「名詞の」としている．また，候補中の「水に」も同様に自立語「水」を品詞 (名詞) に置き換えて，「名詞に」としている．そして，これらの組み合わせ，「入力に「名詞の」を含み，かつ候補に「名詞に」を含む」を特徴量として用いている．入力と候補に含まれる他の全ての 2-gram に対しても同じ処理を行うため，組み合わせの数は「入力中の 2-gram の数 × 候補中の 2-gram の数」個である．

(2) の文節の数では，文節の数は，1 個，2 個，3 個，4 個，5 個以上の 5 種類の特徴量を用いた．**図 4.1** では，「/」が文節を表しており，入力の文節の数は 5 個，候補の文節の数は 6 個である．入力，候補ともに文節数 5 以上なので，「入力の文節数 5 以上かつ候補の文節数 5 以上」を特徴量として用いている．

(3) の同一の自立語の種類と数では，「落ち着く」のように，入力と発話候補に同一の自立語を含む場合，自立語の品詞毎に出現回数を数え，0 個，1 個，2 個，3 個以上の 4 種類の特徴とした．**図 4.1** では，入力と候補の両方に出現する同一の自立語は「落ち着く」という動詞 1 個なので，「入力と候補に同一の動詞を 1 個含む」を特徴量として用いている．

(4) の共起する自立語の種類と数も (3) と同様に．自立語の品詞毎に出現回数を数え，0 個，1 個，2 個，3 個以上の 4 種類の特徴とした．**図 4.1** では，入力と候補の両方に出現する共起する自立語は，「水族館」と「水」，「静か」と「水」，「静か」と「心」，「雰囲気」と「心」という名詞 4 組なので，「入力と候補に共起する名詞を 3 個以上含む」を特徴量として用いている．

本実験では，教師データ内に出現した回数が 100 回未満の特徴を除き，3638 次元の特徴ベクトルを用いた ($n = 3638$)．

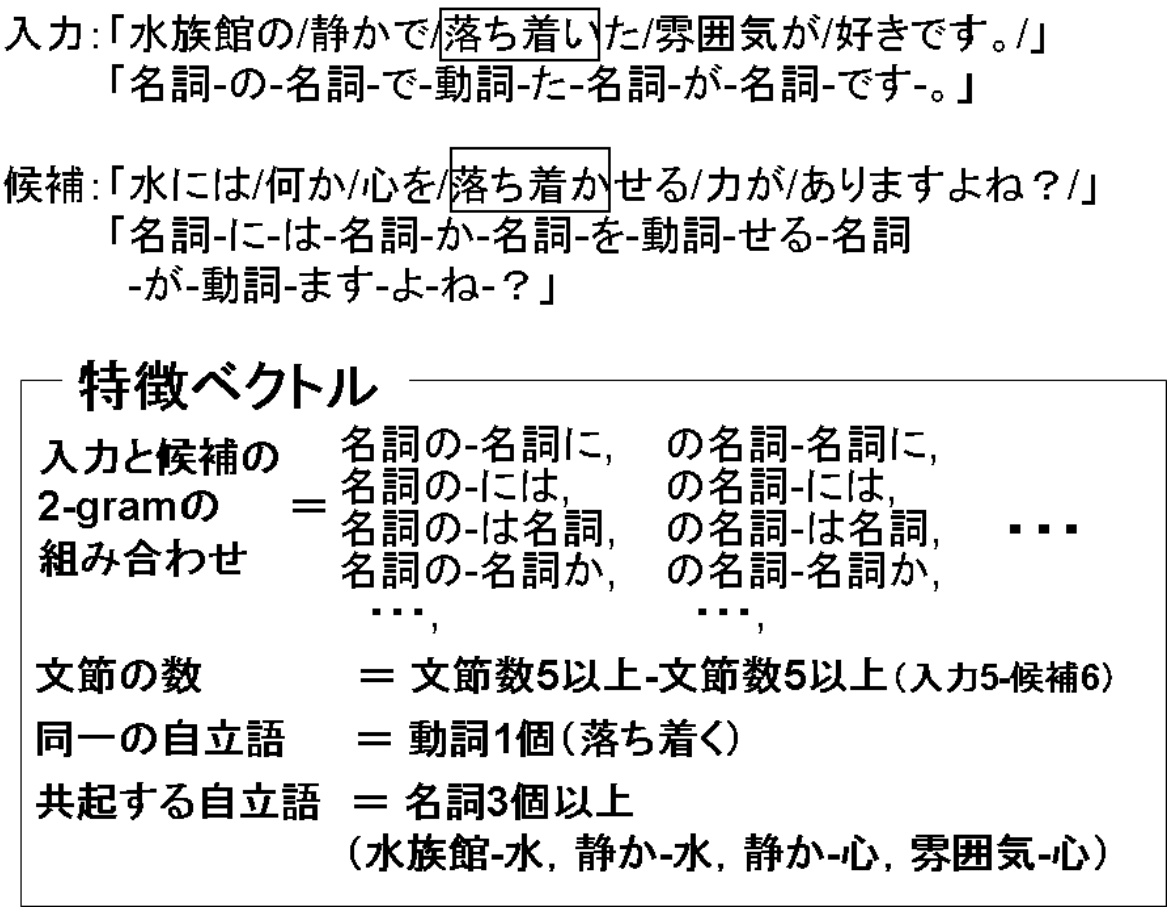

図 4.1　実験で使用した特徴ベクトル

4.4.2　共起情報の抽出

本実験では，特徴ベクトルとして共起する自立語を用いる．共起する自立語を調べることにより，意味的な繋がりを持つ発話候補を選択しやすくなると考えられる．

本研究では，意味的な関連の強さを記述するものとして，Web 日本語 N グラム [73] から抽出した単語の共起情報を用いた [74]．Web 日本語 N グラムは Google によって約 200 億文の日本語データから作成された語彙数が 256 万語の n-gram データであり，1 から 7 の単語 n-gram が収録されている．単語間の共起頻度には，7-gram の中に同時に出現する頻度を用いた．

Web 日本語 N グラムの中に含まれる単語は，単語間の出現頻度の差が非常に大きいため，共起関係の強さを測る指標には，LLS(Log-Likelihood Score)[75] を用いた．ここで，単語 a と単語 b の LLS とは，a と b が互いに従属関係にある場合と，独立関係にあ

表 4.3 「水族館」と共起する単語の例

番号	名称	LLS
1	海	1347152
2	動物	1042871
3	海水浴	698701
4	博物館	642141
5	観光	471946
6	テーマパーク	194108
7	イルカ	144445
8	ペンギン	80175
9	入場	28964
10	水	14436

る場合の尤度比であり，以下の式で表される．

$$LLS = 2 \sum_{i=1,j=1}^{2} f_{ij} \left(log \frac{f_{ij}}{F} - log \frac{f_{i.} f_{.j}}{F^2} \right) \tag{4.18}$$

ただし，$g(a,b)$ を単語 a，b の共起頻度，$h(a)$ を単語 a の出現頻度，F を全文数とすると，$f_{11} = g(a,b)$，$f_{12} = h(a) - g(a,b)$，$f_{21} = h(b) - g(a,b)$，$f_{22} = F - f_{11} - f_{12} - f_{21}$ である．また，$f_{i.} = f_{i1} + f_{i2}$，$f_{.j} = f_{1j} + f_{2j}$ である．単語 a と b の従属関係が強い程，LLS は大きな値を取る．

この LLS は，単語間の出現頻度に差が大きい場合でも，共起関係を検出可能な指標であるといわれている．本実験では LLS を共起関係の強さとし，$LLS > 10000$ の単語ペアを共起関係のある単語ペアとした．例えば，「水族館」と「海」は $LLS = 1347152$ であり，共起している．一方，「水族館」と「俳優」では，$LLS = 625$ であるため，共起していない．「水族館」と共起するその他の単語を**表 4.3** に示す．

4.4.3 実験で用いるデータ

本研究では，Wizard of Oz 法により教師データ C_s を作成した．実験では，大学生の被験者 11 人による対話から 59 発話を抽出し，教師およびテストデータとして用いた．

被験者はあらかじめ休日の過ごし方，将来の夢，自己紹介文などをプロフィールとして自由に記述する．このプロフィールを基に，発話候補集合を被験者別に 11 人の人手で作成した．作成者は被験者の応答を想定し，対話が盛り上がるように注意して発話候補を作

成した．また，発話候補を質問に限定せず，被験者からの質問に対する返答として使用可能な発話や，相槌などの一般的な応答も作成した．

例えば**表 4.2** の a_1^s, a_4^s は，以下に示すプロフィール（一部）から作成した発話候補である．

> 好きな場所は水族館です。薄暗くて静かな落ち着いた感じが好きで昔は月に一回ほどのペースで行っていました。最近は全然行ってないので、余裕ができたら行きたいです。

被験者は，発話候補集合を持つ実験者と対話する．実験者は可能限り発話候補集合から発話を選択して返答を返す．ただし，発話候補集合に適切な応答が含まれない場合，実験者が発話を考えて返答を返す．このようにして取得した対話の中から，実験者が発話候補集合に含まれる応答をしたとき，その一つ前の被験者の発話までを「入力」とする．

各データにおける被験者の入力数，人手で作成した発話候補の数と適切な発話候補数の平均を**表 4.4** に示す．

発話候補数は最低 70，最大で 180 であり，1 つの入力に対する発話候補数の平均は 122 である．入力は，教師データとする対話の 1 状態を示し，各入力の最後の発話と発話候補の組に対して，人手により適切，不適切を評価した．ここで，適切な応答とは，話題を変えず，会話がつながっている応答とした．主観による評価の影響を減らすため，1 つの組につき 3 人で評価し，2 人以上が適切と評価した組を，正解事例，それ以外を誤り事例とした．人手による評価の結果，入力毎の適切な発話候補数の平均は 2.62 となった．

実験では，教師データ C_s として入力 s と，それに対する正解事例 a_c^s を用いた．

発話選択法の性能評価には 59 分割交差検証法を用いた．すなわち，手動で適切，不適切を評価した 59 組のうち，58 組をパラメータ推定のための教師データ，残り 1 組をテストデータとみなして発話候補を順位付け，手動による評価と比較した．テストデータを変更し，59 通り実験を行い，手動で適切と評価された組の順位を確認することにより，発話選択法の性能を評価した．

4.4.4　実験結果

シグモイド関数を用いる手法による順位付けの結果を**図 4.2**，最大エントロピー法を用いる手法の結果を**図 4.3** に示す．

横軸は，適切な発話候補が初めて出現する順位を表し，縦軸はその累積である．すなわち，「その順位以内に適切な応答を最低一つ含む」入力の割合を示している．

図より，最大エントロピー法では，適切な発話候補が 1 位となる入力は 42.4%，発話候補中上位 10 位以内に適切な応答が含まれる入力は 81.4% となることがわかる．同様に，

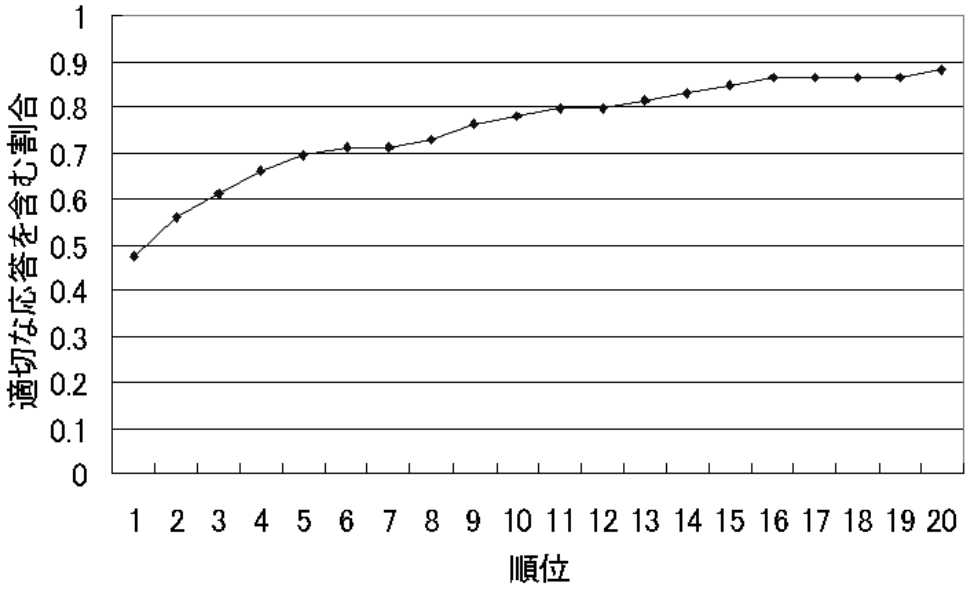

図 4.2　適切な発話候補の順位（シグモイド関数を用いる手法）

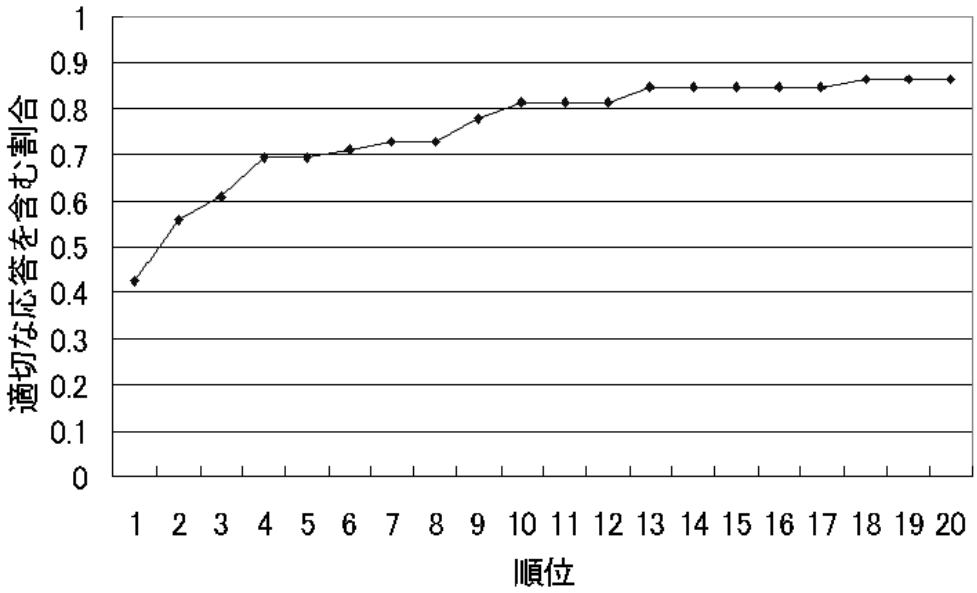

図 4.3　適切な発話候補の順位（最大エントロピー法を用いる手法）

表 4.4　実験で用いるデータ

被験者番号	入力数	発話候補数	適切な発話候補数の平均
1	2	124	7.00
2	6	180	5.00
3	4	70	2.75
4	8	107	2.78
5	10	146	2.08
6	5	161	2.17
7	4	127	4.00
8	6	70	2.67
9	10	130	1.09
10	2	73	1.00
11	2	71	1.33

シグモイド関数を用いる手法では 1 位に 47.5%，10 位以内では 78.0% となることがわかる．

この結果より，どちらの手法も充分高い精度とは言えないが，4 割以上の応答が正しく 1 位となることを確認できた．一方で，上位 10 位までに正しい応答を含む割合は，手法毎の差は小さく，8 割程度と高い精度であることも確認できた．

実際の順位付けの例として，「お菓子は作るのよりも食べるのがすきなんです。」に対する発話候補を，各手法で評価した際の上位 5 位を**表 4.5** に示す．この例では，発話候補の数は 146 個，そのうち適切な応答は 6 個である．表中の発話候補では，「そうでしたか。」，「良く食べますか？」，「どうしてですか？」は適切な応答であり，それ以外の発話候補は不適切であると評価されている．

この例の場合，最大エントロピー法では 1 位，2 位，5 位に適切な応答が含まれており，シグモイド関数を用いる手法も同様に 1 位，2 位，5 位に適切な応答が含まれている．このことから両手法とも，発話候補の順位付けに有効であることが確認できる．

本手法を詳細に分析するため，シグモイド関数を用いる手法による学習後の $\hat{\boldsymbol{w}}$ を調べる．重要な特徴は $|w|$ の値が大きくなると考えられる．実際，$-1.0 \leq w \leq 1.0$ と値が小さい特徴が大半であり，3299 個と全体の 90.7% を占めている．それ以外の，$|w| > 1.0$ となった重要と考えられる特徴の一部を**表 4.6** に示す．この表から，同じ語や，共起する語

表 4.5　発話候補の順位付けの例

手法	順位	発話候補
シグモイド関数を用いる手法	1	良く食べますか？
	2	そうでしたか。
	3	いえいえ，ご謙遜を。
	4	「好きこそものの上手なれ」と言いますから。
	5	どうしてですか？
最大エントロピー法	1	そうでしたか。
	2	良く食べますか？
	3	いえいえ，ご謙遜を。
	4	「好きこそものの上手なれ」と言いますから。
	5	どうしてですか？

表 4.6　学習後の $\hat{\boldsymbol{w}}$

番号	特徴	w
1	同じ名詞を 1 つ含む	3.28
2	共起する語を 2 つ含む	2.01
3	共起する形容詞を 1 つ含む	1.85
4	共起する動詞を 2 つ含む	1.64
5	同じ動詞を 1 つ含む	1.12
6	文節数 2-文節数 4 以上	-1.65
7	NOUNwo-文節数 2	-2.05
8	共通する語を含まない	-3.02

が含まれると選択されやすくなるように w を学習したことが確認できる．すなわち，共起する語や，同一の語を含む発話は対話において重要であるといえる．

4.5　要約

本章では，機械学習アルゴリズムを用いた，対話エージェントのための統計的な発話選択法を提案した．発話選択を，入力に対して大量に用意された発話候補集合の中から，応答として適切な発話を選択する問題と定義した．発話候補を評価するため，発話候補を人手で評価したデータを学習データとし，相対的な大小関係を学習した．手法の有効性を確認するため，人手により適切とされた応答の順位を確認した．実験の結果，入力に対して適切な発話候補が 1 位となる入力はシグモイド関数を用いる手法では 47.5%，最大エントロピー法を用いる手法では 42.4% であった．また，発話候補中上位 10 位以内に適切な応答が含まれる入力は，それぞれ 78.0%，81.4% となることを確認した．

対話エージェントに本手法を実装する場合は，適切な応答が 1 位となることが重要であり，どちらの手法もまだ改善の余地がみられる．

一方，人間の応答を補助するエージェントに本手法を実装する場合は，エージェントの提案する候補に適切な応答が含まれることが重要である．そのため，上位 10 位以内に適切な応答が 78.0% と多く含まれている本手法を用いることにより，このようなエージェントを実現できる可能性が確認された．

今後の課題としては，大量の発話集の自動生成があげられる．本論文では，対話相手のプロフィールに基づいた大量の発話候補が事前に用意されているものとして実験しているが，発話候補集は自動生成することが望ましい．今回，発話候補集の生成は事前に人手で作成するスタティックな生成法を用いているが，対話の流れに応じて発話候補集を自動生成するダイナミックな生成法に対しても本手法は適用可能である．その場合，対話の流れなど，発話候補集の生成に必要な特徴量を追加し，さらに特徴ベクトルから発話候補集を自動生成する処理も追加する必要がある．

さらに，教師データの量と，発話自動選択法の精度の関係を調査し，学習に必要となる教師データのサイズを求めることも今後の課題である．

第 5 章

結論

5.1 本論文のまとめ

本研究では，非タスク指向型対話エージェントの客観的・定量的な評価法と，発話の自動選択法を提案した．

第 2 章では，非タスク指向型対話エージェントを客観的・定量的に評価する技術について述べた．対話エージェントを評価するため，対話エージェントの行った対話の自然さを評価した．対話の自然さの評価法として，人間同士の対話との類似度を用いた．前提として，人間同士の対話を理想的な対話であると仮定し，人間同士の対話と人間と非タスク指向型対話エージェントとの対話との類似度を計算することにより，対話の自然さを評価した．まず, 対話に, 談話の浅い構造を表現できる Switchboard Discourse Annotation and Markup System of Labeling(SWBD-DAMSL) タグをまとめたものである簡易 DAMSL タグの系列を付与し，タグの系列で表現した．次に，類似度の計算のため，隠れマルコフモデル (Hidden Markov Model(HMM)) によって人間同士の対話タグ系列をモデル化した．評価すべき対話をこの HMM に入力し，HMM の出力確率を計算することにより，対話の自然さを評価した．

対話の自然さの評価実験では，人間による主観評価と HMM による評価を比較した．その結果，人間に近い自然な対話を行う対話エージェントほど出力確率が高くなることを確認した．さらに，複数の対話エージェントを評価することにより，人間同士の対話は人間と対話エージェントとの対話より HMM の出力確率が高くなり，その中でも自然な対話ほど出力確率が高くなることを確認した．これによって，HMM の出力確率の大小によって対話の自然さを評価できることが明らかになった．

ただし，本実験では HMM による評価法の有効性のみを確認するため，タグ付与を全て手動で行い，タグ付与による誤りはないものとしている．しかし，実際に本手法を適用する場合には，全ての処理を自動化する必要があるため，タグ付与は自動化する必要がある．

第 3 章では，第 2 章の手法を発展させ有用性を高めるため，一連の処理のうち，手動タグ付与の部分を Conditional Random Fields(CRF) によって自動化する方法について述べた．まず，手動でタグを付与した標準データを用いて，CRF のパラメータを学習した．CRF を用いて各発話に自動タグ付与を施し，全ての対話をタグ系列として表現した．そして，人間同士の対話のタグ系列を学習した HMM を作成した．

次に，自動タグ付与の有効性を実験により確認した．実験の結果，自動タグ付与の正解率は 75.77%，意味的に近いものを許容した場合で 83.20% という結果を得た．kappa 統計量は 0.6371 であった．以上のことから，CRF を用いた自動タグ付与の有効性が確認できた．

次に，自動付与したタグの系列を学習した HMM による対話の評価実験を行った．実験の結果，より人間に近い自然な対話を行う対話エージェントほど HMM の出力確率が高くなることを確認した．さらに，人間同士の対話は人間と対話エージェントとの対話より HMM の出力確率が高くなり，その中でも自然な対話ほど出力確率が高くなることを確認した．この結果から，自動タグ付与を用いた HMM は，手動タグ付与を用いた HMM と同等の性能を実現できることが明らかとなった．これにより，提案した評価法の全処理を自動化し，評価法としての有用性を確認できた．

第 4 章では機械学習アルゴリズムを用いた，対話エージェントのための統計的な発話選択法について述べた．発話選択を，入力に対して大量に用意された発話候補集合の中から，応答として適切な発話を選択する問題と定義した．発話候補を自動評価するために，入力と発話候補の対に対して，「適切」または「不適切」を手動で評価し，教師データとした．教師データから発話候補の評価値の相対的な大小関係を学習し，発話選択を行った．本手法の有効性を評価するために，発話候補に対する手動評価の結果と自動評価の結果を比較した．実験の結果，入力発話に対して適切な応答が 1 位となる割合は 47.5%，発話候補中上位 10 位以内に適切な応答が含まれる割合は 78.0% となることを確認した．ただし，対話エージェントに本手法を実装する場合は，適切な応答が 1 位となることが重要であり，本手法にはまだ改善の余地がみられる．

一方，人間の応答を補助するエージェントに本手法を実装する場合は，エージェントの提案する候補に適切な応答が含まれることが重要である．そのため，上位 10 位以内に適切な応答が 78.0% と多く含まれている本手法を用いることにより，このようなエージェントを実現できる可能性が確認できた．

5.2 今後の課題

5.2.1 対話エージェントの評価法に関する今後の課題

対話エージェントの評価において，今後検討すべき課題としては以下が挙げられる．

対話の表現方法

本研究では，対話を表現するために簡易 DAMSL タグ系列を用いた．しかしながら，十分な量の標準データを利用できるのであれば，SWBD-DAMSL タグを使うことで評価の精度が上がる可能性がある．さらに，DAMSL タグ以外のタグを使用することで，深い構造や，発話間の関係などを表現することが可能である．例えば，徳久らの研究 [76] では，対話を表現するためのタグとして，Dialog Act(DA) タグと，Rhetorical Relation(RR) タグを独自に定義し，これらのタグと対話の盛り上がりとの関連を調べている．これらのタグのうち，DA タグは発話の行為を示し，RR タグは，表層の意味関係と深層の意味関係の両方を記述することが可能である [77, 78]．徳久らは，SWBD-DAMSL タグと MRDA タグ [79] を組み合わせ，47 種類の DA タグを定義している．さらに，表層の意味関係に着目して 16 種類の RR タグを定義している．

本研究では，評価の計算に HMM を使用するため，系列データとして扱うことが可能なタグであれば，タグの種類によらず類似度を計算することが可能である．そこで，上記の RR タグを用いることで，対話の自然さ以外の評価も可能になると考えられる．

自動タグ付与の精度向上

提案手法の性能をより向上させる方法として，簡易 DAMSL タグの自動付与手法の改良が考えられる．たとえば，標準データの量を増やすことで，自動付与の精度を上げることができる．また，タグを付与したい対話に対して，内容の似ている標準データを用いることによって精度を上げることができると考えられる．

本研究では識別モデルである CRF を使用してタグ付与を行った．しかし，近年では，生成モデルが再評価されつつあり，特に，ノンパラメトリックな手法が研究され，急速に発展してきている [65][66]．この手法を応用することで，クラスタ数を指定することなく，発話をクラスタリングすることが可能である．そこで，各クラスタをタグと捉えることで，あらかじめタグの種類を指定することなく自動タグ付与が可能となる．

発話と対話を表現する特徴ベクトルの変更

自動タグ付与では，タグ付与の対象となる発話を特徴ベクトルで表現する．本研究では 1-gram,2-gram と一つ前の発話のタグを特徴ベクトルとして使用した．上記以外の特徴量として，発話の自動選択で用いた共起情報，品詞，話題の焦点，発話間の依存関係などを使用できると考えられる．ただし，特徴量の種類を増やすと，タグ付与に必要な標準データの量も増大するため注意が必要である．

自然さ以外の評価

本手法では，対話の自然さを評価の基準としているが，非タスク指向型対話エージェントを設計するためには，他の視点からの評価も必要である．たとえば，人間を楽しませることを目的とした対話エージェントを設計する際には，自然さも必要であるが，対話の楽しさも評価する必要がある．このような場合でも，適切な入力系列と HMM を用いることによって評価することが可能であると期待される．

さらに，今後，非タスク指向型対話エージェントの性能が向上した場合には，本論文で提案したような，発話間の繋がりの自然さを評価するだけでは不十分である．話者の意図や目的，発話の焦点，主題等の「深い構造」に着目した評価法に発展させる必要がある．

5.2.2　対話エージェントの設計に関する今後の課題

対話エージェントの設計において，今後検討すべき課題は数多い．その中で，特に本研究に関係する課題を挙げる．

大量の発話集の自動生成

本論文では，対話相手のプロフィールに基づいた大量の発話候補が事前に用意されているものとして実験しているが，発話候補集は自動生成することが望ましい．例えば，対話エージェント ELIZA[3] は，発話候補をテンプレートに基づいて生成している．本研究で目標とするインタビュア型対話エージェントでは，対話相手のプロフィールに基づいた大量の発話候補を生成する必要がある．雑談のような非タスク指向型対話のための発話候補生成法として，Web 検索の結果を用いる手法 [80] が提案されている．しかし，適切な応答として利用可能な割合は 6 割程度であることや，人間がメインテーマを選定する必要がある半自動生成であることなど，発展の余地は大きい．

自動選択手法の精度向上

自動選択において，教師データの種類・量と機械学習の手法の選択が重要である．非タスク指向型対話エージェントの評価において発話を表現する特徴ベクトルが重要であったように，自動選択においても入力と発話候補の組を表現する特徴ベクトルが重要である．特に，自然言語処理の分野では特徴ベクトルの次元数が数万以上と大きくなりがちなため，次元の呪い [81] に注意しなければならない．例えば，工藤らの係りうけ解析 [69] や，Web 検索の順位付け [82] など，単語，URL，またはそれらの組み合わせを特徴としているため，特徴数が 2 万以上となっている．これらの研究では，SVM の使用や，パラメータの正則化などの対策を行っている．また，URL と query の cosine 距離や，単語の意味など，複雑な特徴を利用することで精度が上がる可能性がある．さらには，教師データの量と，発話自動選択法の精度の関係を調査し，学習に必要となる教師データのサイズを求めることも今後の課題である．

謝辞

本研究は，研究室所属当時からの石井健一郎教授による懇切丁寧な御指導によるものであります．教授には本研究を進めるにあたり，終始温かく見守って頂き心より感謝いたします．また，本研究について，多岐に渡り御指導いただいた鳥海不二夫助教に厚く御礼申し上げます．

また，本論文をまとめるにあたり，貴重な御意見をいただいた名古屋大学大学院情報科学研究科社会システム情報学専攻の渡邉豊英教授，間瀬健二教授，および松原茂樹准教授に心より感謝いたします．

同様に，様々な形でご協力頂いた同研究室所属の藤田幸久氏，稲葉通将氏，河田純奈氏，光崎夏紀氏，山口竜一氏，太田知宏氏，中村大介氏，西岡寛兼氏，田島裕之氏に心より感謝の意を表します．また，発話候補自動選択法における評価実験を手伝って頂いた，同研究室所属の太田健文氏，神谷達幸氏，平井尚樹氏に心より感謝いたします．

また，刺激的な議論と多くの示唆を頂いた同研究室 OB の曽我部将義氏，村田早織氏，小澤猛志氏，李相浩氏，久連石圭氏，小泉康治氏，岡田譲二氏，菊地安仁氏，原大曜氏，石田健氏に心より感謝いたします．

本研究の実施にあたり，NTT コミュニケーション科学基礎研究所の前田英作氏，および南泰浩氏に御意見を頂きました．記して感謝いたします．

そして，対話実験および手動タグ付与にご協力いただいたインターグループ様，対話実験に参加してくださった皆様に心より感謝いたします．

参考文献

[1] 厚生労働省. 平成 20 年国民生活基礎調査の概況, http://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa08/dl/gaikyou.pdf. 2009.

[2] 藤田雅博. ロボットエンターテイメントと人工知能 (特集「エンターテイメントとai」). 人工知能学会誌, Vol. 16, No. 3, pp. 399–405, 2001.

[3] J. Weizenbaum. ELIZA-a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, Vol. 9, No. 1, pp. 36–45, 1966.

[4] T. Winograd. *Understanding Natural Language*. Academic Press, Inc. Orlando, FL, USA, 1972.

[5] V. Lesser, R. Fennell, L. Erman, and D. Reddy. Organization of the Hearsay II speech understanding system. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 23, No. 1, pp. 11–24, 1975.

[6] V. Zue, J. Glass, D. Goodine, H. Leung, M. Phillips, J. Polifroni, and S. Seneff. Integration of Speech Recognition and Natural Language Processing in the MIT VOYAGER System1. *Proc.IEEE-ICASSP*, Vol. 1, pp. 713–716, 1991.

[7] J.R. Glass and T.J. Hazen. Telephone-based conversational speech recognition in the jupiter domain. In *Fifth International Conference on Spoken Language Processing*. Citeseer, 1998.

[8] S. Seneff and J. Polifroni. Dialogue management in the Mercury flight reservation system. In *Proc. ANLP-NAACL*, pp. 1–6, 2000.

[9] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, and V. Zue. Galaxy-II: A reference architecture for conversational system development. In *Fifth International Conference on Spoken Language Processing*. Citeseer, 1998.

[10] J. Polifroni and S. Seneff. Galaxy-II as an architecture for spoken dialogue evaluation. In *Proceedings of the Second International Conference on Language Re-*

sources and Evaluation. Citeseer, 2000.

[11] P.R. Cohen, A. Cheyer, M. Wang, and S.C. Baeg. An open agent architecture. In *AAAI Spring Symposium*, Vol. 18, 1994.

[12] A. Cheyer and L. Julia. Multimodal maps: An agent-based approach. *Lecture Notes in Computer Science*, Vol. 1374, pp. 111–121, 1998.

[13] S. Seneff. Interactive computer aids for acquiring proficiency in Mandarin. *Lecture Notes in Computer Science*, Vol. 4274, p. 1, 2006.

[14] W. Burgard, A.B. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun. Experiences with an interactive museum tour-guide robot. *Artificial Intelligence*, Vol. 114, No. 1-2, pp. 3–55, 1999.

[15] R. Siegwart, K.O. Arras, S. Bouabdallah, D. Burnier, G. Froidevaux, X. Greppin, B. Jensen, A. Lorotte, L. Mayor, M. Meisser, et al. Robox at Expo. 02: A large-scale installation of personal robots. *Robotics and Autonomous Systems*, Vol. 42, No. 3-4, pp. 203–222, 2003.

[16] N. Dahlbäck, A. Jönsson, and L. Ahrenberg. Wizard of Oz studies: why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces*, p. 200. ACM, 1993.

[17] S. Dow, B. MacIntyre, J. Lee, C. Oezbek, J.D. Bolter, and M. Gandy. Wizard of Oz support throughout an iterative design process. *IEEE Pervasive Computing*, pp. 18–26, 2005.

[18] 駒谷和範, 上野晋一, 河原達也, 奥乃博. ユーザモデルを導入したバス運行情報案内システムの実験的評価. 情報処理学会研究報告. SLP, 音声言語情報処理, Vol. 2003, No. 75, pp. 59–64, 2003.

[19] 駒谷和範, 鹿島博晶, 田中克明, 河原達也. 複合的言語制約に基づくキーフレーズ検出を用いた汎用的なデータベース検索音声対話プラットフォーム. 情報処理学会論文誌, Vol. 44, No. 5, pp. 1333–1342, 2003.

[20] 翠輝久, 駒谷和範, 清田陽司, 河原達也, 木戸冬子. 音声対話による大規模知識ベース検索システム: 音声版ダイアログナビ (音・音声インタフェース). 情報処理学会研究報告. SLP, 音声言語情報処理, Vol. 2004, No. 74, pp. 21–26, 2004.

[21] M. Nakano, Y. Minami, S. Seneff, T.J. Hazen, D.S. Cyphers, J. Glass, J. Polifroni, and V. Zue. Mokusei: A telephone-based Japanese conversational system in the weather domain. In *Seventh European Conference on Speech Communication and Technology*. Citeseer, 2001.

[22] Junichi Fukumoto, Tsuneaki Kato, Fumito Masui, and Tsunenori Mori. An overview of the 4th question answering challenge (qac-4) at ntcir workshop 6. In

Noriko Kando and David Kirk Evans, editors, *Proceedings of the Sixth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, pp. 433–440, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan, May 2007. National Institute of Informatics.

[23] R. Higashinaka and H. Isozaki. NTT ' s question answering system for NTCIR-6 QAC-4. In *Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access (NTCIR-6)*, pp. 460–463, 2007.

[24] T. Sakai, Y. Saito, Y. Ichimura, M. Koyama, and T. Kokubu. Toshiba ASKMi at NTCIR-4 QAC2. *NTCIR-4 QAC2 Working Notes*, 2004.

[25] H. Isozaki, K. Sudoh, and H. Tsukada. NTT ' s japanese-english cross-language question answering system. In *Proceedings of the NTCIR Workshop 5 Meeting*, pp. 186–193, 2005.

[26] C. Hori, T. Hori, H. Tsukada, H. Isozaki, Y. Sasaki, and E. Maeda. Spoken interactive ODQA system: SPIQA. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 2*, pp. 153–156. Association for Computational Linguistics Morristown, NJ, USA, 2003.

[27] R. Higashinaka, K. Dohsaka, S. Amano, and H. Isozaki. Effects of Quiz-style Information Presentation on User Understanding. In *Proc. of Interspeech*, pp. 2725–2728, 2007.

[28] R. Speer, J. Krishnamurthy, C. Havasi, D. Smith, H. Lieberman, and K. Arnold. An interface for targeted collection of common sense knowledge using a mixture model. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*, pp. 137–146. ACM, 2009.

[29] http://www.w3.org/tr/voicexml21/.

[30] E. Reiter and R. Dale. Building applied natural language generation systems. *Natural Language Engineering*, Vol. 3, No. 01, pp. 57–87, 1997.

[31] J.R. Glass, J. Polifroni, and S. Seneff. Multilingual language generation across multiple domains. In *Third International Conference on Spoken Language Processing*. ISCA, 1994.

[32] A. Stent, R. Prasad, and M. Walker. Trainable sentence planning for complex information presentation in spoken dialog systems. In *Proceedings of ACL*, 2004.

[33] F. Mairesse and M. Walker. Learning to personalize spoken generation for dialogue systems. In *Ninth European Conference on Speech Communication and*

Technology. Citeseer, 2005.

[34] R.E. Schapire. A brief introduction to boosting. In *International Joint Conference on Artificial Intelligence*, Vol. 16, pp. 1401–1406. LAWRENCE ERLBAUM ASSOCIATES LTD, 1999.

[35] A.H. Oh and A.I. Rudnicky. Stochastic language generation for spoken dialogue systems. In *ANLP/NAACL 2000 Workshop on Conversational systems-Volume 3*, p. 32. Association for Computational Linguistics, 2000.

[36] B.J. Grosz and C.L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, Vol. 12, No. 3, pp. 175–204, 1986.

[37] M.F. McTear. Software to support research and development of spoken dialogue systems. In *Sixth European Conference on Speech Communication and Technology*. Citeseer, 1999.

[38] E. Levin, R. Pieraccini, and W. Eckert. A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on Speech and Audio Processing*, Vol. 8, No. 1, pp. 11–23, 2000.

[39] S. Singh, M. Kearns, D. Litman, and M. Walker. Reinforcement learning for spoken dialogue systems. In *Proc. NIPS99*. Citeseer, 1999.

[40] M. Minsky. A framework for representing knowledge. *In The Psychology of Computer Vision*, 1974.

[41] R.W. Smith, A.W. Biermann, and D.R. Hipp. An architecture for voice dialog systems based on prolog-style theorem proving. *Computational Linguistics*, Vol. 21, No. 3, pp. 281–320, 1995.

[42] D. Jurafsky and J.H. Martin. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. MIT Press, 2000.

[43] R. Soricut and E. Brill. A Unified Framework for Automatic Evaluation using N-gram Co-Occurrence Statistics. *Proceedings of ACL*, pp. 613–620, 2004.

[44] Litman D.J. Kamm C.A. Walker, M.A. and A. Abella. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics*, pp. 271–280, 1997.

[45] 木村泰知, 荒木健治, 桃内佳雄, 栃内香次. 遺伝的アルゴリズムを用いた帰納的学習による音声対話処理手法. 電子情報通信学会論文誌, Vol. J84-D2, No. 9, pp. 2079–2091, 2001.

[46] AM TURING. Computing machinery and intelligence-AM Turing. *MIND*, Vol. 59, p. 236, 1950.

[47] Toriumi F. Sogabe, M. and K. Ishii. An evaluation method of non-task-oriented dialog systems. *IPSJ SIG Technical Report*, 2005. in Japanese only.

[48] Shriberg E. Jurafsky, D. and D. Biasca. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual. *Institute of Cognitive Science Technical Report*, pp. 1–61, 1997.

[49] H.P. Grice. *Studies in the way of words*. Harvard University Press Cambridge, Mass, 1991.

[50] 曽我部将義, 小澤猛志, 石井健一郎. テキスト対話における対話戦略とその評価法. 電気関係学会東海支部連合大会予稿集,O-414, 2004.

[51] Toriumi F. Okada, J. and K. Ishii. Automatic generation of question sentence for computerized non-task-oriented dialogue agent that imitated interviewer. *Proceedings of JAWS 2007*, 2007. in Japanese only.

[52] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257–286, 1989.

[53] Rabiner L.R. Levinson, S.E. and M.M. Sondhi. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *The Bell System technical journal*, Vol. 62, No. 4, pp. 1035–1074, 1983.

[54] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, Vol. 7, No. 2, pp. 179–188, 1936.

[55] N. Reithinger and M. Klesen. Dialogue act classification using language models. *Proceedings of EuroSpeech-97*, pp. 2235–2238, 1997.

[56] http://chasen.naist.jp/hiki/ChaSen/.

[57] C.J. van Rijsbergen. *Information Retrieval: Second Edition*. Butterworths, London, 1979.

[58] Sanchis E. Castro M.J. Grau, S. and D. Vilar. Dialogue Act Classification using a Bayesian Approach. In *9th Conference Speech and Computer*. ISCA, 2004.

[59] Guyon I.M. Boser, B.E. and V.N. Vapnik. A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on computational learning theory*, pp. 144–152, 1992.

[60] U. Kressel, et al. Pairwise classification and support vector machines. *Advances in Kernel Methods–Support Vector Learning*, pp. 255–268, 1999.

[61] McCallum A. Lafferty, J.D. and F.C.N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the Eighteenth International Conference on Machine Learning (table of contents)*,

pp. 282–289, 2001.

[62] R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge Univ Pr, 1998.

[63] C.D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, 2002.

[64] F. Sha and F. Pereira. Shallow Parsing with Conditional Random Fields. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp. 213–220, 2003.

[65] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[66] C. Kemp, J.B. Tenenbaum, T.L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *Proceedings of the National Conference on Artificial Intelligence*, Vol. 21, p. 381. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[67] 春野雅彦, 白井諭, 大山芳史. 決定木を用いた日本語係受け解析. 情報処理学会論文誌, Vol. 39, No. 12, pp. 3177–3186, 1998.

[68] 内元清貴, 関根聡, 井佐原均. 最大エントロピー法に基づくモデルを用いた日本語係り受け解析. 情報処理学会論文誌, Vol. 40, No. 9, pp. 3397–3407, 1999.

[69] 工藤拓, 松本裕治. 相対的な係りやすさを考慮した日本語係り受け解析モデル. 情報処理学会論文誌, Vol. 46, No. 4, pp. 1082–1092, 2005.

[70] S. Della Pietra, V. Della Pietra, J. Lafferty, R. Technol, and S. Brook. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 4, pp. 380–393, 1997.

[71] D.C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, Vol. 45, No. 1, pp. 503–528, 1989.

[72] M.F. Møller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, Vol. 6, pp. 525–525, 1993.

[73] 工藤拓, 賀沢秀人. Web 日本語 n グラム第 1 版.

[74] 稲葉通将, 磯村直樹, 鳥海不二夫, 石井健一郎. 意味ネットワークによる非タスク指向型対話システムの評価. 電子情報通信学会技術研究報告, Vol. 108, No. 456, pp. 29–34, 2009.

[75] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, Vol. 19, No. 1, pp. 61–74, 1993.

[76] 徳久良子, 寺嶌立太. 雑談における発話のやりとりと盛り上がりとの関連. 人工知能学会論文誌, Vol. 21, No. 2, pp. 133–142, 2006.

[77] W.C. Mann and S.A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, Vol. 8, No. 3, pp. 243–281, 1988.

[78] A. Stent and J. Allen. Annotating argumentation acts in spoken dialog. Technical report, University of Rochester, 2000.

[79] R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg. Meeting recorder project: Dialog act labeling guide. *ICSI, Berkeley, CA, USA, Tech. Rep. TR-04-002*, 2004.

[80] 柴田雅博, 冨浦洋一, 西口友美. 雑談自由対話を実現するための WWW 上の文書からの妥当な候補文選択手法. 人工知能学会論文誌, Vol. 24, No. 6, pp. 507–519, 2009.

[81] 石井健一郎, 上田修功, 前田英作, 村瀬洋. わかりやすいパターン認識.

[82] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 133–142. ACM New York, NY, USA, 2002.

発表文献リスト

論文

(1) 磯村直樹, 鳥海不二夫, 石井健一郎. HMM による非タスク指向型対話システムの評価. **電子情報通信学会論文誌**, Vol.J92-D, No.4, pp.542-551, 2009.

(2) 磯村直樹, 鳥海不二夫, 石井健一郎. 対話エージェント評価におけるタグ付与の自動化. **電子情報通信学会論文誌**, Vol.J92-A, No.11, pp.795-805, 2009.

国際会議・ワークショップ

(1) N. Isomura, F. Toriumi, K. Ishii. Statistical Utterance Selection Using Word Co-occurrence for a Dialogue Agent. *PRIMA 2009*, LNAI 5925, pp.68-79, 2009.

国内研究会・ワークショップ

(1) 磯村直樹, 鳥海不二夫, 石井健一郎. 対話エージェントのための統計的発話候補選択法. **人工知能学会第** *86* **回 知識ベースシステム研究会**, pp.33-38, 2009.

(2) 磯村直樹, 鳥海不二夫, 石井健一郎. インタビュア型対話エージェントの設計. **人工知能学会** *MYCOM2009(***第** *10* **回** *AI* **若手の集い***)*, pp.95-98, 2009.

(3) 稲葉通将, 磯村直樹, 鳥海不二夫, 石井健一郎. 意味ネットワークによる非タスク指向型対話システムの評価. **電子情報通信学会技術研究報告**, Vol.108, No.456, pp.29-34, 2009.

(4) 磯村直樹, 鳥海不二夫, 石井健一郎. HMM による非タスク指向型対話エージェントの対話構造のモデル化. *Human-Agent Interaction Symposium 2008 (HAI-2008)*, pp.1-6, 2008.

(5) 磯村直樹, 鳥海不二夫, 石井健一郎. 対話エージェントにおける非タスク指向型対話評価法の提案. *Joint Agent Workshops and Symposium 2008(JAWS2008)*,

pp.1-8, 2008.

(6) 磯村直樹, 鳥海不二夫, 石井健一郎. HMM による非タスク指向型対話システムの性能の比較評価. **電子情報通信学会技術研究報告**, Vol.107, no.246, pp.1-6, 2007.

(7) 磯村直樹, 鳥海不二夫, 石井健一郎. HMM による非タスク指向型対話システムの評価. **電子情報通信学会技術研究報告**, vol.106, no.298, pp.39-44, 2006.