学　士　論　文

題　目　心療カウンセリングにおける会話データのカテゴリ分類に関する研究

指導教員　小山田 耕二 教授

京都大学工学部 電気電子工学科

氏 名　林田祐磨

平成 29 年 2 月 10 日

目 次

第1章 序論

心療において, ストレス等に依る心の悩みを持ったクライエントに対してカウンセラーはカウンセリングを行う. その中でカウンセラーがクライエントの問題意識, つまりどのような「対人関係上の問題」を感じているかに注意して会話を引き出すことがカウンセリングの基本である [1].

しかし新人カウンセラーは熟練カウンセラーに比べて, 自らの質問に対するクライエントの回答として,「対人関係上の問題」を引き出す事に関して未熟であるという問題がある. クライエントから「対人関係上の問題」を引き出すためには, カウンセラーの質問内容が重要とされている.

そこで新人カウンセラーに対し, 熟練カウンセラーが指導を行うスーパービジョンの機会が設けられている. スーパービジョンの事例として,「心療内科における摂食障害専門ヨーガ療法グループ」事例検討会 [2] では新人カウンセラーとクライエントのカウンセリング内容を動画で撮影し, その動画から書き起こされたテキストデータを熟練カウンセラーが読んで議論を行っている. その中で, 新人カウンセラーは, 自分の関心でカウンセリングを進めてしまいがちで, 自分の中で作り上げた解釈内容をクライエントに確認するための「閉じられた質問」を多用する傾向が顕著であるとされる. しかし熟練カウンセラーがカウンセリング内容の文字を読むだけで, これらの会話の流れを十分に理解することは困難とされている. そのため, カウンセラーのどのような質問に対して, クライエントからどのような回答が得られたかについて, 視覚的に理解できる仕組みが求められていた.

上辻らはこのような要求に対して, カウンセリングにおける会話の流れの可視化システムを開発した. それによって, このシステムではカウンセリングの会話の流れが時間軸に沿って可視化され, カウンセラーのどのような質問で, クライエントのどのような回答を引き出せたかについて, 視覚的に理解できることを示した [3].

会話の流れの可視化システムの中でカウンセリングテキストデータとして取り扱うヨーガ療法では, アドラー心理学が取り入れられている. アドラー心理学では,「人生のすべての問題は3つの主要な課題に分類することができる.

つまり, 交友の課題, 仕事の課題, 愛の課題である.」と唱えた[4]. 以上より, ヨーガ療法ではクライエントの課題を, クライエントにとっての親疎の関係から,

- 仕事の課題 : 永続しない人間関係
- 交友の課題 : 永続するが, 運命を共にしない人間関係
- 愛の課題 : 永続し運命を共にする関係

の 3 つに区別している. また, カウンセリングにおいて, クライエントが発言した話題がどの課題に関するものか, さらにカウンセリングの会話のやりとりの中でその課題に関する発言がどのように推移しているかを分析することは, カウンセリングプロセスを明確にするために役立つ. したがって, 会話の流れの可視化システムでは, クライエントの会話内容を上で述べた 3 つの課題について, カテゴリ分類を行っている.

会話の流れの可視化システムではクライエントの発言内容を 1 発言単位ではなく 1 文単位で「愛」,「交友」,「仕事」に分類している. ここで 1 文とは, 句点やクエスチョンマークで区切られた単位である. また, 会話の流れ可視化システムにおけるクライエントの発言内容の分類手法は以下の通りである.

- クライエントの発言を書き起こしたテキストデータの各文に対して, 形態素解析を行い, 文を各単語に分ける
- 各カテゴリに含まれると思われる単語をカテゴリ毎に指定しておき, クライエントの発言文を構成する単語と一致する単語の数が一番多いカテゴリに分類し, カテゴリ毎に指定された単語と 1 つも一致しなかった文は「未分類」とする

しかし上に述べた分類手法では, 分類結果が予め指定した単語に大きく依存するといった問題がある. 例えば,「友達」と「友だち」のような表記ゆれや,「夫」と「旦那」などの同義語が, 同じ意味の単語であると認識できないといったことである. さらに,「夫の仕事がいつも遅いのが原因であまり眠れません.」という文は,「愛」のカテゴリに分類されるべきであるが,「仕事」という単語が「仕事」カテゴリの辞書に登録されている「仕事」カテゴリに分類されてしまうという問題もある. このような誤った分類結果に関しては, ユーザーである熟練カウンセラーがシステム上で分類結果を確認し, 誤った分類がなされ

ている箇所を手動修正する必要があり, 作業の負担がかかる問題が指摘されている. そのため, カウンセリングにおけるクライエントの発言内容をより正確に自動分類することが求められている.

本研究では, 機械学習を用いてクライエントの発言内容に対するカテゴリ分類を行い, 会話の流れの可視化システムにおける分類手法と比べて高い分類精度で自動カテゴリ分類を行うことができるか検証を行う.

本論文の構成は次の通りである. 第 1 章は本論文の序論である. 第 2 章では, 本論文の関連研究を挙げる. また本研究での分類手法に関する関連知識について説明する. 第 3 章では, 本研究における, クライエントの発言内容の分類手法について述べ, 第 4 章では本研究における分類結果を述べ, 本研究における分類結果と会話の流れ可視化システムにおける分類結果との比較に対しての考察について述べる. 第 5 章では本論文の結論と, 本研究の今後の課題について述べる.

第2章 関連研究

本章では, 本研究との関連研究を示し, その位置づけについて述べる.

テキスト分類は, スパムメールの自動振り分けやニュース記事の自動分類など様々な目的で行われている. 近年, 機械学習を用いたテキスト分類に関する研究が盛んに行われ, その中でも教師付き学習によるテキスト分類に関する研究が数多く行われている.

平ら[5)]は機械学習によるテキスト分類問題に対して, 出現頻度の小さい単語まで考慮した学習を行わなければ分類精度が落ちることを述べ, 高次元の単語ベクトルを用いるために SVM を用いた学習を行うことで, ニュース記事の分類において高い分類精度を実現した. この研究ではニュース記事の各文書の中から名詞を抽出し,Bag-of-words モデルによって文書をベクトル化している. ここで Bag-of-words とは文書中の単語の並びなどは考えず, 文書に単語が含まれているかどうかのみを考えるモデルである.

従来ではニュース記事のように, ある程度文章が長く単語数や単語の種類も多い文書のベクトル化の際には Bag-of-words モデルがよく使われてきたが, 本研究での分類対象はクライエントの発言 1 文毎であり, 文の長さが短く単語数も少ないものが多い. そのため名詞を抽出し Bag-of-words モデルを用いて各文をベクトル化しても, その文の特徴が表れにくいと考えられる.

短い文の分類では,Sriram ら[6)]は Twitter の Tweet 内容を「ニュース」や「イベント」など 5 つの目的別に自動カテゴリ分類する手法を提案した. この研究では Bag-of-words による素性に加えて,Tweet の中に略語やスラングが使われているか, 時間や場所についての記載があるか, など 8 つの特徴を基にした素性も加えることで,Bag-of-words のみによる素性を用いるよりも高い精度で Tweet の自動カテゴリ分類を行った.Sriram らの研究での分類カテゴリは, 例えば「イベント」であれば時間や場所についての記載が多い, などのカテゴリ毎の Tweet の特徴が顕著であるが, 本研究での分類カテゴリである「愛」「仕事」「交友」にはそのような顕著な特徴は存在しない.

また,Bag-of-words モデルのデメリットとして,「友達」と「友だち」などの表記ゆれや,「父親」と「父さん」のような同義語を, 全く別の単語として捉える

といった点が挙げられる. さらに, 文書のベクトル次元数が学習する全コーパス中の語彙数と等しくなるため, 本研究でも学習コーパスの語彙数は約 2 万でありベクトル次元数も 2 万にも及ぶ. そのためニューラルネットワークを用いて学習を行うと, 計算時間が膨大になるという問題がある Bag-of-words によるベクトルを次元圧縮したものを用いる手法も考えられるが, 永田ら [7] によって次元圧縮により分類精度が下がったことが示されている.

Mikolov ら [8] は単語の分散表現を学習して単語のベクトル化を行う word2vec を提案した.word2vec では数百次元程度の密なベクトルで単語を高い精度で表現することが可能であり, 現在もその用途について様々な研究が行われている.

word2vec により得られた単語の分散表現を用いて, 単語間の意味的な類似度を求めることが可能であり, 日本語の研究として単語の意味を取り扱う研究が行われている. 野沢ら [9] は, 大量のレシピデータから食材と調理法を抽出し,word2vec で学習させ,word2vec で得られた単語ベクトルから各単語に類似する単語を算出し代替食材を発見する手法を提案した. また, 菅原ら [10] は単語の分散表現を用いて多義語の語義曖昧性を解消する手法を提案した. 語義曖昧性というのは, 例えば”cool”という単語は「涼しい」や「かっこいい」など複数の語義を持つために, 文脈により語義が異なることを言う. そこで word2vec により得た単語の分散表現を用いて, 文書中における多義語の最もふさわしい語義を選ぶことを目的としている.

しかし,word2vec により得た単語ベクトルを基に文章ベクトルを作成し, 機械学習による文章分類に応用している日本語の研究事例は少ない.

日本語以外を取り扱う,word2vec により得た単語ベクトルを基にした文書分類では,Xing ら [11] は,word2vec で得た単語ベクトルと,LDA モデルを用いた単語ベクトルを用いてそれぞれで文書ベクトルを作成した後に機械学習を行い, 中国語のニュース記事の自動分類精度を比較した. 機械学習アルゴリズムとしてはナイーブベイズ,k 近傍法,SVM を用い, その結果 word2vec で得た単語ベクトルを基に文書ベクトルを作成し SVM で機械学習を行う分類手法が最も精度が高かったことを示した.

また加藤ら [12] は商品に対するレビューデータと評点に対し,word2vec と深層学習を用いて評判分析を行い,1-of-K ベクトルを用いたロジスティック回帰の性能とほぼ同程度であることを示した. しかし先に述べたように, 本研究ではクライエントの発言 1 文毎を分類対象としているため,1-of-K ベクトルを用い

ての分類は不適当と考える.

本研究における手法として,word2vec により単語の分散表現を学習して得た単語ベクトルを基に, 知恵袋の悩み相談に関する質問文を 1 文毎にベクトル化し,SVM とニューラルネットワークによる機械学習を行い, クライエントの発言 1 文毎に対する自動カテゴリ分類を行った. その後, 本研究における分類手法と会話の流れ可視化システムにおける分類手法との比較を行った.

第3章 分類手法

本章では, クライエントの発言内容のカテゴリ分類の手法について詳しく述べる. なお本研究では, 会話の流れの可視化システムと同様にクライエントの発言内容を1文毎に分類している. クライエントの1文とは, 句点かクエスチョンマークで区切られた単位のことである. また, クライエントの発言内容は上でも述べたように「愛」「交友」「仕事」という3つのカテゴリに分類する.

3.1 分類手法の概要

本研究における分類手法の概要を図 3.1 に示す. まず,Yahoo!知恵袋 [13] の悩み相談に関する文章をコーパスとして,word2vec を用いて, 単語のベクトルを得る.word2vec により得た単語のベクトルを基に,3つのどのカテゴリに属しているかのラベルを持った Yahoo!知恵袋の悩み相談に関する文のベクトルを教師付きデータとする. 同様の手順で, クライエントの発言内容を書き起こしたテキストデータを1文毎にベクトル化し, これに正解カテゴリを付与したものをテストデータとする. 教師付きデータを機械学習の入力データとし, 学習させて, テストデータを学習したモデルに入力し, 出力された予測カテゴリと予め与えられている正解ラベルとが一致するか調べ評価を行う. 3.2 節から 3.5 節にかけて本研究での分類における各段階の詳細について述べる.

3.2 word2vec を用いた単語のベクトル化

本研究で扱うデータは様々な単語から構成される文章の集まった文書データである. しかしコンピュータで単語や文を扱うために, 単語や文を数値ベクトルとして表現する必要がある. 機械学習により自動でテキストデータの分類を行う研究はこれまでに様々行われてきたが, 多くの機械学習のアルゴリズムでは入力の次元数を学習前に定める必要がある. 扱うテキストデータは単語の数, 文や文章の長さがそれぞれ異なるものがほとんどであり, 入力するテキストデータを固定長のベクトルで表現する必要がある. このような課題に対して

これまでに主に用いられている手法は bag-of-words である.bag-of-words は文章を単語の集合として捉え, 単語の並び方などは考慮せずに, 単語の出現の有無と出現回数のみを考慮する手法である. しかしこの手法を用いると, 学習に用いるテキストデータに含まれる語彙数だけベクトルの次元数が増えてしまい, 語彙数と次元数が等しくなる. そのためニューラルネットワークなどのネットワークを用いた機械学習での入力次元が数万～数十万程度の次元になり, 学習を行う際に計算時間が膨大になる. このような次元数の大きいベクトルに対して特異値分解による次元圧縮を行い, それを用いた分類 [6] も行われているが, 数千次元まで圧縮すると分類精度が落ちることが報告されている. そこで, 単語の分散表現を学習することで単語を数百次元程度の低次元のベクトルで表現する手法である word2vec が Mikolov ら [7] により提案された.bag-of-words のようにベクトルの 1 つの次元のみが値を持ち, その他の次元の値が 0 となる疎なベクトル表現とは違い, 分散表現では全ての次元が値を持っている. 本研究においても機械学習を行う際に扱うテキストデータには数万以上の語彙数が含まれているため word2vec を用いて単語のベクトル化を行った.

word2vec に学習させるコーパスとしては Wikipedia やニュース記事のような大規模データが用いられることが多い. そこから得られた単語の分散表現を用いて機械学習などを用いて分類を行う研究では, 分類の対象となるのがニュース記事などであることが多いためである. しかし本研究で扱うテキストデータはカウンセリングデータであり会話文である.Wikipedia やニュース記事は, 固有名詞が比較的多く出現するが, 個人の感情を表す形容詞はあまり含まれていない. しかしカウンセリングデータは会話文であるため固有名詞は少なく, 感情を表す形容詞は比較的多く含まれる. しかし, カウンセリングの内容を書き起こしたテキストデータはあまり多く存在しておらず, 且つプライベートな内容であるためインターネットなどでは公開されておらず, 収集するのが困難である. そこで, カウンセリングデータとより似た内容の文章をコーパスに用いることを考えた. そのためカウンセリングデータに比較的内容が近いと思われる Yahoo!知恵袋に投稿されたものの中で悩み相談に関するカテゴリから文章を取得した. 具体的には, 本研究で分類を行う「愛」「交友」「仕事」の 3 カテゴリと内容が類似していると思われる以下の 4 カテゴリから取得した. 4 カテゴリの中から,1 つの質問とそれに対してのベストアンサーを 1 件として, 各カテゴリ 12000 件, 計 48000 件を取得してコーパスとした.

- 愛：　「恋愛相談」,「家族関係の悩み」
- 交友：　「友人関係の悩み」
- 仕事：　「職場の悩み」

また,word2vec は元々英語の文章を対象に考えられているため, コーパスは単語同士が半角スペースで区切られている必要がある. そのため, 本研究では形態素解析ツール MeCab14) を用いて予め知恵袋の文章を単語毎に区切った文章を入力した. また, コーパスに対して形態素解析を行う際に, 全ての単語を基本形で出力している.

次に word2vec で学習を行う際のパラメータについて述べる. ベクトルの次元数に関して,word2vec の開発者である Mikolov らによると, コーパスのデータサイズが増えるにつれてベクトルの次元数も大きくすべきとされている 15). また, データサイズがあまり大きくないにも関わらずベクトルの次元数を大きくしすぎると, 精度が落ちることも報告されている. デフォルトの次元数は 100 次元であり, 本研究で用いたコーパスはデータサイズがおよそ 40MB, 語彙数がおよそ 1 万語と大きくないため, ベクトルの次元数は 200 以下とした. 窓長は各窓長に対して学習後の各単語の意味的な類似度を調べた際に適切な値とされる 5 に設定した. 学習アルゴリズムとして cbow モデルか Skip-gram モデルを選択できるが, 本研究では Skip-gram モデルを選択した. また,min-count の値は 5 とし, コーパス内に 5 回未満しか出現しない単語は考慮しないとした. このような手順で単語の分散表現を得た.

3.3　文のベクトル化

次に 3.2 節で述べた単語の分散表現を用いて, 文をベクトルに変換する.word2vec により単語の分散表現を得た際と同様にして, まず文を単語毎に区切る必要がある. そして文中に出現する各単語のベクトルを word2vec により得た分散表現から見つけ出し, それを用いて文ベクトル化を行う. 具体例として,「母ともよく喧嘩しますし.」という文ベクトル化の流れを図 3.2 に示す.

- 「母ともよく喧嘩しますし.」という文に対して形態素解析により単語毎の分かち書きを行い, 基本形で出力する.

- 文中に出現する全ての単語に対して, 各単語のベクトルを word2vec により学習したモデルから取り出す. ただし文末の句点とクエスチョンマークは除いた.
- 「母」,「とも」,「よく」,「喧嘩」,「する」,「ます」,「し」の単語ベクトルの 1 次元目の値を全て足し合わせ, 各単語のベクトルの 1 次元目の和を計算する. このとき, もし単語が word2vec により学習したモデルに存在せずベクトルが存在しなかった場合, その単語は無視する.
- 1 次元目に対して行った手順をベクトルの次元全てに対して行い, 全ての次元分の和を計算する
- 各文の長さは異なり, また単語の数も異なるので, 足し合わせた単語数で各次元の和を割り, これを文のベクトルとする

つまり, 各単語と文ベクトルの次元数を M として, 文ベクトルを

$$d = [d_1, d_2, d_3, ?, d_M]$$

とし, 足し合わせた単語の,i 番目の単語の j 次元目の要素の値を

$$w_i j$$

とすると,

$$d_1 = \frac{1}{N}\sum_{i=1}^{n} w_i 1, d_2 = \frac{1}{N}\sum_{i=1}^{n} w_i 2, ?, d_M = \frac{1}{N}\sum_{i=1}^{n} w_i M$$

と表せる.Xing ら 10) は, 機械学習による文書分類を行う際に,word2vec で得た単語ベクトルの単純な和の平均として文章ベクトルを作成する手法により, 分類精度が高くなることを示している. したがって本研究でも, 文ベクトルを単語ベクトルの和の平均として算出した.

3.4 文ベクトルを入力とする機械学習

本節では, 本研究における学習と分類手法について述べる. 本研究の目的は, 予め指定された 3 つの「愛」「交友」「仕事」カテゴリに, どのカテゴリに属す

るかが未知である文を正確に自動カテゴリ分類することである. 本研究で行う学習は教師付き学習であり, 教師付き学習とは, 入力のデータに対して出力が指定されて学習を行う方法である. 本研究では,3.3 節で述べた手順で算出した 1 文毎の文ベクトルを, 入力である 1 文毎の特徴量として用いて, 出力には「愛」「交友」「仕事」の 3 つの分類カテゴリを用いる. 学習を行ったパターン認識器に対して, どのカテゴリに属するか未知であるクライエントの発言 1 文毎の文ベクトルを入力し, 予測カテゴリを出力させることで, 自動カテゴリ分類を行う.

本研究では, 機械学習に用いるアルゴリズムに SVM16) とニューラルネットワーク 17) を用いた.

SVM とはパターン認識モデルの 1 つであり, 線形しきい素子を用いて 2 つのカテゴリのパターン識別機を構成する手法である.2 つのカテゴリに分類を行う際に, ハードマージン SVM では入力データが完全に線形分離可能であると仮定して,1 つの誤分類も許容せずに分離超平面を決定する. つまり,M 個の m 次元教師付きデータ xi(i=1, 2,⋯, M) がクラス 1, 2 いずれかに属するとして, ラベルをクラス 1 のときに yi=1, クラス 2 のときに-1 とする. 線形分離可能である場合, 決定関数は w を m 次元係数ベクトル,b はバイアス項として

$$D(x) = w^T x + b$$

と決めることが出来る. また線形分離の条件から決定関数は

$$y_i(w^T x_i + b)?1(i = 1, ?, M)$$

の条件を満たす. 条件式 (4) を満たすような w,b を求め, 入力データに対して分離超平面を決定するが, このとき分離超平面とそれにもっとも近い教師データとの間の距離をマージンとよぶ. 条件式 (4) を満たす分離超平面は無数に存在するが,SVM ではマージンが最大となる超平面を識別境界とする. したがって,SVM では汎化能力が高いというのが大きな特徴である.

しかし, 本研究では入力データは完全に線形分離可能ではないと考え, ソフトマージン SVM を用いた. ソフトマージン SVM では線形分離可能でない場合に適用できるように,

$$\xi_i(?0)$$

を導入し条件式 (4) を拡張し,

$$y_i(w^T x_i + b)?1 - \xi_i, (i = 1, ?, M)$$

を満たし,

$$Q(w, b, \xi) = \frac{1}{2} \| w \|^2 + C \sum i = 1^M \xi_i$$

を最小にするような最適化問題を考え分離超平面を決定する. したがってハードマージン SVM とは異なり, ソフトマージン SVM では誤分類を許容する. ここで C は式 (6) の右辺第 1 項のマージン最大化と第 2 項の誤分類の最小化のトレードオフを決定するパラメータであり, 分離超平面の決定に大きな影響を与えるため, 適切に設定する必要がある.　そして, このように入力空間内で非線形分離を行う必要がある場合には入力空間を高次元の特徴空間に写像して特徴空間上でマージンが最大となるように超平面を決定する方式がよく用いられており, カーネル法とよばれているが, 本研究でもこの方式を用いた. カーネル法では, 非負のラグランジュ乗数

$$\alpha_i, \beta_i$$

を導入して, 決定関数は

$$D(x) = \sum i \in \ S \alpha_i y_i K(x_i, x_j) + b$$

となる. ここで

$$K(x_i, x_j)$$

はカーネルであり, 問題に応じてこのカーネルを適切に選ぶことで汎化能力を向上することができる. 本研究では,

$$K(x, x') = x^T x'$$

で表される線形カーネル (linear) と,

$$K(x, x') = exp(-\gamma \| x - x' \|^2)$$

(9) で表されるラジアル基底関数カーネル (rbf) を用いた. ここで γ は分布の半径を制御するパラメータであり,rbf カーネルにおいて超平面の決定に大きな影響を与える.

また, 上でも述べたように SVM は 2 つのカテゴリに分類する 2 値分類を基本としているが, 本研究では 3 つのカテゴリに分類することを目的としている.SVM において多カテゴリへの分類を行う場合,One-Vs-One 方式と One-Vs-The-Rest 方式のいずれかを用いる. 今,K カテゴリに分類することを考える.One-Vs-One 方式はある特定のカテゴリに属するか, また別の特定のカテゴリに属するかの 2 カテゴリ分類問題を解く分類器を, 全てのカテゴリの組み合わせ, つまり K(K-1)/2 個使用する.One-Vs-The-Rest 方式はある特定のカテゴリに属するか, 他の K-1 個のいずれかのカテゴリに属するかの 2 カテゴリ分類問題を解く分類器を, 全ての組み合わせ,K 個利用する.

ニューラルネットワークもパターン認識に用いられるモデルの 1 つである. 本研究で用いたニューラルネットワークは入力層と出力層を 1 つずつ持ち, 中間層を複数持つ多層ニューラルネットワークである. また全てのノード同士がリンクによって繋がれた, 全連結型のネットワークである. ニューラルネットワークの仕組みの概要として, 入力層の各ノードに値を入力し, それぞれの値が重みを持ったノード間のリンクを通って, 閾値を持った中間層の各ノードに受け渡され, その値が中間層で変換される. 中間層で変換される際には活性化関数を用いて変換が行われる. そして変換された値が同じように次の中間層へと受け渡されていき, 最後に出力層に伝達し値を出力させる. そして出力した値が期待される値に近づくように, 各ノード間のリンクの重みやノードの閾値を最適化するように学習を行う.

つまり, 本研究では入力層のノード数を各文ベクトルの次元数と等しく固定し, ベクトルの各要素の値を入力層の各ノードに入力する. 出力層のノード数は分類カテゴリ数と同じ 3 で固定し, 出力が最も大きいカテゴリをその文が属するカテゴリであると判断する. また中間層の層数, ノード数は可変なパラメータとしており, このパラメータによって精度が変わるため,SVM と同様に後述の実験により最適なパラメータを求めた.

本研究での分類の手順としては, 初めに 3.3 節で作成した教師付きデータを上で述べた機械学習の入力として学習を行う. そして学習後のモデルにテストデータを入力し,「愛」「交友」「仕事」のいずれかの予測カテゴリを出力させ 3 カテゴリへの分類を行う.

以上の分類法をもとに，本研究では，時間軸に沿った可視化によって，来談者と治療者の会話の流れを可視化し，治療者の発した質問によって会話の流

れがどのように変化するのかを明らかにするWebシステムを開発した．次章では本論文の提案システムに対する要件の抽出，および設計と実装について説明する．

第4章 実験内容,結果と考察

本章では，ユーザーである熟練治療者からのコメントをもとにした提案システム要件抽出，および本提案システムの設計や実装について詳しく述べる．なお本システムでは，来談者または治療者の1発言を，話者の交代によって判断するものとする．つまり，来談者の1発言は，治療者が喋り終わって来談者が喋り始めてから，来談者が喋り終わって治療者が喋り始めるまでである．また，治療者の1発言は，来談者が喋り終わってから，あるいは治療者が話を切り出し始めてから，治療者が喋り終わって来談者が喋り始めるまでである．

4.1 システム要件

まず専門家にインタビューを行い，次に述べるシステム要件を抽出した．

本提案システムの要件は大きく2点挙げられる．1つ目の要件は，カウンセリングにおける来談者と新人治療者との会話の流れの可視化である．カウンセリングの会話の流れのテキストデータは治療者からの発話によってまったく展開が異なってくるので，実際の来談者個人の症状の分析よりも，治療者トレーニングとして利用するのに適しているからである．本提案システムは，治療者の能力向上に資する可視化分析システムに関するものとして，治療者トレーニングの客観的指標は臨床心理学領域で期待がされている．

2つ目の要件は，来談者と治療者のそれぞれの発言を分類である．来談者の発言としては，どの1文がどの課題領域（仕事，交友，愛）に属するのか，治療者の発言としては，どの発言がオープン（5W1Hを問うもの）またはクローズ（Yes/Noを問うもの）なのかを分類表示できることが必要である．来談者からの提出した話題が，どの領域に関するものか，カウンセリングの中でその領域がどのように変わっていくかを分析していくと，そのカウンセリングプロセスがより明確になると考えられる．時系列に沿ってカウンセリングの会話の流れを見ることで，最初にクライエントの関心がどこにあったのか，それに対して治療者からの発話で異なった領域に話題が展開した，というような分析も可能になると考えられる．来談者側の関心に注意を向けると

いうことがカウンセリングの基本であるため，治療者側の関心で来談者を誘導してしまうことはよくないとされる．そのためのチェックが本システムによって可能になることが期待される．もうひとつの課題としては，新人治療者は閉じられた質問を多用するので，新人治療者が開かれた質問を自由に使えるように熟練治療者が指導する必要がある．そのチェックも本システムによって可能になることが期待される．

来談者の発言に関する分類

来談者の発言に関する分類に関しては，1発言単位ではなく1文単位で「愛」「交友」「仕事」を分類するように変更することが求められる．たとえば,「うちの夫は仕事にいくのを嫌がって，毎朝起きてこないんです．それを見ているだけで腹が立つんです．自分の同僚が同じように仕事に行きたがらなくて朝起きなかったという話を聞いても，さほど腹は立ちませんが，夫がそうなるのは絶対に許せない」という文章について，専門家は次の通りに分類する．第1文は愛の課題，第2文は仕事の課題，第3文は愛の課題に分類する．原文から，治療者から来談者への質問事項の分類の指標として，来談者の発言をうながして来談者自身も気づいていなかったことを認知させるのが大事であるので，それぞれの発言量の可視化の実装を盛り込むべきであるというコメントを得た．ここから得られる要件は，来談者の発言の分量に応じて，それをはさむ治療者の発言の縦棒をグラフ上で変えることであると考えた．

治療者の発言に関する分類

治療者の発言の分類について，まず治療者の発言も，次章で述べる初期分類は適当ではないことがあるので，来談者からの返答だけでなく，治療者の質問区分けも手動で修正したいというコメントを得た．また，治療者の発言について,「開かれた質問」「閉じられた質問」だけでなく,「解釈」「相槌」「無駄話」という分類を追加してほしいというコメントを得た．来談者の発言に関する分類とは異なり，治療者の発言に関する分類は1文単位ではなく1発言単位での分類でよいというコメントを得た．

4.2　システム設計と実装

前節では提案システムのユーザーである熟練治療者からのコメントをもとにした，提案システム要件抽出について説明した．本節では提案システムの設計と実装について説明する．このシステムに模擬会話データを入力した際の描画結果のスクリーンショットを図 4.1 に示す．本節ではその要件をみたした提案システムの設計と実装について説明する．なお，本提案システムの開発言語は JavaScript である．本提案システムの構成を図 4.2 に示す．

まず，グラフ描画前のテキストデータ処理について述べる．アクティビティ図を図 4.3 に示す．アクティビティ図とはフローチャートに似た図で，いわゆるビジネスロジックにおける手続き的な流れやプログラムの制御フローを表す UML の図である．UML は，主にオブジェクト指向分析や設計のための，記法の統一がはかられた（Unified）モデリング言語（Modeling Language）である．

初めに，ブラウザ上で各テキストの会話データを読み込んで，テキストを単語ごとに区切る形態素解析を行う．会話データ読み込み前の本提案システムスクリーンショットを図 4.4 に示す．形態素解析サブシステムには，JavaScript 言語の形態素解析ライブラリである kuromoji.js[14] を使用した．kuromoji.js は，Java で実装されたオープンソースの日本語形態素解析エンジン Kuromoji[?] を，JavaScript に移植したものである．

形態素解析された単語群から，句点やクエスチョンマークを終点と定義して 1 文ずつのテーブルをつくる．さらに，話者の切り替わりを全角コロンで定義し，全角コロンと全角コロンの間の文のグループを 1 発言と定義し，発言単位でテーブルをつくる.「来談者の発言においてこれを含む文はこのグループに属するだろう」という単語を,「愛」「仕事」「交友」ごとに指定しておき，発言データの文を構成する単語が指定された単語と一致する場合，この情報を図 4.2 の発言データ分類サブシステムに引き渡し，発言データ分類の初期設定値を計算する．

質問内容は，前章の要件通り 5 種類に分類され，後述する縦棒アイコンの色の割り当ては次の通りである．濃いピンク色は 5W1H「いつ」「どこで」「誰が」「何を」「どのように」「どうした」などで問われるような「開かれた質問」，濃い青色は Yes/No で答えられる，あるいは一言だけで簡単答えられるような「閉じられた質問」，紫は「相槌」，オレンジは来談者の問題を治療者

がどう「解釈」しているかの確認，黒は「無駄話」を表現している．第1章で述べた通り，来談者が何に問題意識を感じているかをカウンセリングで引き出すには，治療者は「閉じられた質問」よりも「開かれた質問」をしたほうがよいとされている．

治療者の1発言についても，来談者の発言の1文ごとの分類と同様に，関連する単語をあらかじめ指定しておき，必要情報を 4.2 の発言データ分類サブシステムに引き渡し，発言データ分類の初期設定値を計算する．ここで会話データを入力した後の治療者の初期分類状態の分類方法を簡単に図 4.5 に示す．まず「いつ」「どこ」「何」などのいわゆる 5W1H を示す疑問詞をもつ発言を「開かれた質問」に分類する．次に残りの発言のうち，単語数が 5 個以下のものを「相槌」に分類する．さらに残りの発言のうち終助詞「か」を含む発言を「閉じられた質問」に分類する．今までの 3 つに分類されなかったものは「解釈」か「世間話」に分類されるわけだが，終助詞の「ね」を含むものを「解釈」，含まないものを「世間話」とした．

以上がグラフ描画前のテキストデータ処理である．その後，治療者からの質問形態と，「愛」「仕事」「交友」の文のグループの時間経過に沿った分布変化を可視化する．グラフ描画の際に，会話データを発言データ可視化サブシステムに受け渡す．このサブシステムでは，データ可視化ライブラリ D3.js[15] を使用した．

来談者発言ビュー

図 4.1 において，まず積み重ね折れ線グラフは来談者の1回の発言の中での，アドラー心理学の各カテゴリの分布を可視化している．ただし1回の発言は，話者の交代を発言の区切れ目とする．この積み重ね折れ線グラフにおいて青色は「仕事関係」，ピンク色は「愛（恋愛･愛関係)」，緑色は「交友（友人関係)」に密接に関係する単語を含む文の分布を表している．横軸は時間軸を表現している．ただし対象データである書き起こしテキストデータからは実際の経過秒数は読み取れないので，読み込ませたデータ内における来談者の発言量の累計の文字数を横軸，つまり時間軸とした．後述する治療者の質問を表現する縦棒はこの時間軸にそって出現する．治療者と来談者の発言の入れ替わりがわかりやすいように，積み重ね折れ線グラフが来談者の発言のグループ分布をちょうど表しているのは便宜上発言部分の中央部分になって

いる．縦軸は発言中の 1 文の数のうち，どのグループに何文入っているかという文の数を表している．

治療者発言ビュー

Eric ら[16] は，トピックモデリングによって分けたトピックの時間分布を上下の非対称積み重ね折れ線グラフによって可視化した．一方本研究では，治療者と来談者の会話において，来談者は文ごとに，治療者は発言ごとに描画を行いたい，かつ選択肢表示のためにグラフを省スペースしたいという観点から，治療者の発言を横軸より下に折れ線グラフとして描画するのではなく，来談者の発言のグループ分布を示す積み重ね折れ線グラフに重ねて縦棒として表示するようにした．こうすることによって，来談者と治療者の発言が交互に描画されるので，来談者と治療者の発言の関連性がわかりやすくなった．積み重ね折れ線グラフに重なっている縦棒は，治療者の質問内容の形態を示す．

発言分類手動修正機能

積み重ね折れ線グラフが描画された後に，グラフ左下のラジオボタンエリアにて，クライアントの発言の各 1 文および治療者の各 1 発言において，分類を変えると即時にグラフ描画に変更が適用されるようにした．

原文表示機能

どこがどの発言を表しているかわからないという問題に対しては，治療者から来談者への質問を示す縦棒をマウスオーバーすることによって，グラフ右下に周辺の発言を表示する機能を追加することで解決を図った．見やすさのため，グラフの色に対応した文字色ではなく黒い文字で各発言を表示し，それをグラフと対応した色の隅付き括弧【】で囲うようにした．

第5章 結論と今後の課題

5.1 結論

本研究では,word2vec によって得られた単語のベクトルから 1 文毎の文ベクトルを算出し,SVM を用いて機械学習を行い, クライエントの発言 1 文毎の自動カテゴリ分類を行い, 会話の流れの可視化システムにおける辞書に基づいたカテゴリ分類と比較して, 高い分類精度を実現した.

しかし、前章で述べたとおり, ユーザーの手動修正の負担がかからない分類精度には及んでおらず, 分類精度をさらに高める必要がある. さらにもっと多くの未知のクライエントの発言文に対して分類精度を検証する必要がある. 次節に今後の課題について述べる.

5.2 今後の課題

本研究で得られたことを踏まえて, 今後検討するべき課題について述べる.

- 「あの人の仕事を否定したくありません.」というような文を正しく分類するために, 文の係り受け解析を行い,「あの人」の「仕事」であることから「愛」のカテゴリであると分類できるようにすること
- クライエントの前の発言に遡らないとカテゴリが把握できない文や, カウンセラーからの質問に遡らないとカテゴリが把握できない返事などの短い文を正しく分類するために, クライエントの発言 1 文毎の分類ではなく, カウンセラーの質問やクライエントの前の発言を含めた, 会話のブロック毎の分類
- 主語が抜けている文や,「あの人」のように主語の特定が困難である文を正しく分類するために, その前の会話に遡り主語を補うことができるかの検討
- さらなる未知のクライエントの発言文に対する分類と,「交友」カテゴリに属するクライエントの発言文に対する分類の検討

- 教師付きデータとして,Yahoo!知恵袋の質問文以外のテキストデータの検討
- 学習データ数による分類の正答率の推移の検討
- 分類カテゴリに「自己」,「スピリチュアル」の追加

謝 辞

本研究を進めるにあたり, 有益な御指導, 御助言を頂きました京都大学学術情報メディアセンタービジュアリゼーション研究分野の小山田耕二教授, 江原康生特定准教授, 夏川浩明特定助教, 尾上洋介特定助教に深く感謝致します.

本研究を進めるにあたり, プログラミング技術を始め, 様々な御助言を頂き協力して下さった, 京都大学大学院人間・環境学研究科修士課程 2 年生の今井晨介氏, 京都大学大学院工学研究科修士課程 1 年生の上辻智也氏, 梅澤浩然氏をはじめとする院生の先輩の皆様にはご協力を賜りました. ここに深く御礼申し上げます.

最後に, 家族をはじめとする私の学生生活を支えてくださったすべての皆様へ心から感謝の意を表します.

参 考 文 献

1) 野田俊作, 続アドラー心理学トーキングセミナー勇気づけの家族コミュニケーション, (アニマ 2001, 1991).

2) T. Asano, 認定ヨーガ療法士会, `http://yogatherapy-hyogo.net/lecture.html`, (2015).

3) Atilika, Yahoo!知恵袋, `http://chiebukuro.yahoo.co.jp/`, (2012).

4) H. L. Ansbacher and R. R. Ansbacher, *THE INDIVIDUAL PSYCHOLOGY OF ALFRED ADLER*, (Haper Row Publishers Inc, New York, 1956).

5) 博順平, 隆文向内, 雅彦春野, Support vector machine によるテキスト分類, 情報処理学会研究報告自然言語処理（NL）, Vol. 1998, No. 99, (1998), pp. 173–180.

6) Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas, Short text classification in twitter to improve information filtering, *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, (ACM, New York, NY, USA, 2010), pp. 841–842.

7) 佐々木稔永田純平, 文書分類をタスクとした pylearn2 の maxout+dropout の利用, 言語処理学会第 21 回年次大会, (2015), pp. 900–903.

8) Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, Efficient estimation of word representations in vector space, *CoRR*, Vol. abs/1301.3781, (2013).

9) 健人野沢, 義貴中岡, 修平山本, 哲司佐藤, word2vec を用いた代替食材の発見手法の提案 (データ工学と食メディア), 電子情報通信学会技術研究報告. DE, データ工学, Vol. 114, No. 204, (2014), pp. 41–46.

10) 菅原拓夢, 笹野遼平, 高村大也, 奥村学, 単語の分散表現を用いた語義曖昧性解消, 言語処理学会発表論文集, Vol.21, (2015), (), pp. 648–651.

11) Xuewei Zhang Chao Xing, Dong Wang, Document classification with distributions of word vectors, *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference*, ().

12) 加藤和平, 大島考範, 二宮崇, word2vec と深層学習を用いた大規模評判分析, 言語処理学会第 21 回年次大会発表論文集, (2015), pp. 525–528.

13) Kuromoji morphological analyzer, `http://www.atilika.org`, (2012).

14) T. Asano, kuromoji.js, `https://github.com/takuyaa/kuromoji.js`, (2015).

15) B. Michael, D3. js, *Data Driven Documents*, (2012).

16) E. Alexander and M. Gleicher, Task-driven comparison of topic models, *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS*, Vol. 22, No. 1, (2016), pp. 320–329.

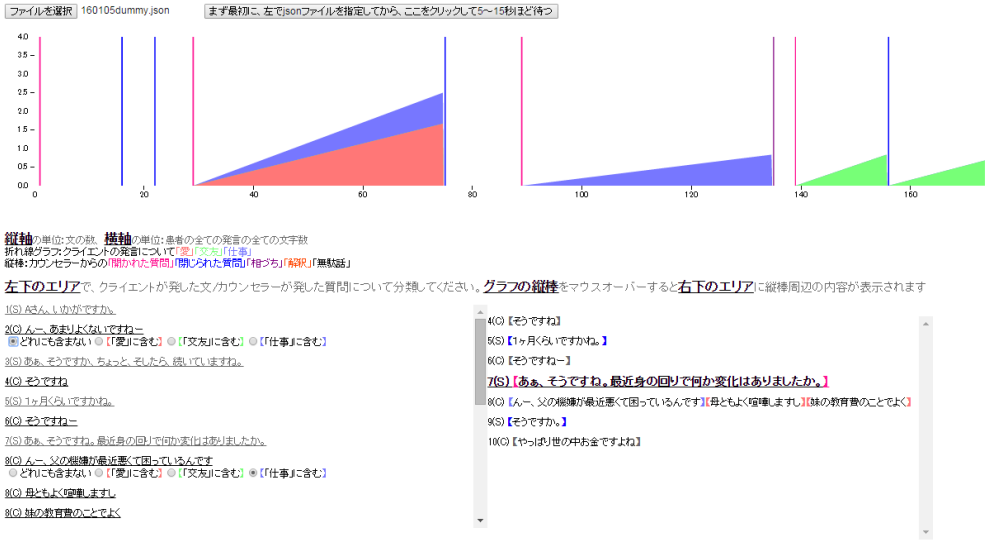

図 4.1: 提案システムの模擬会話データでの可視化結果

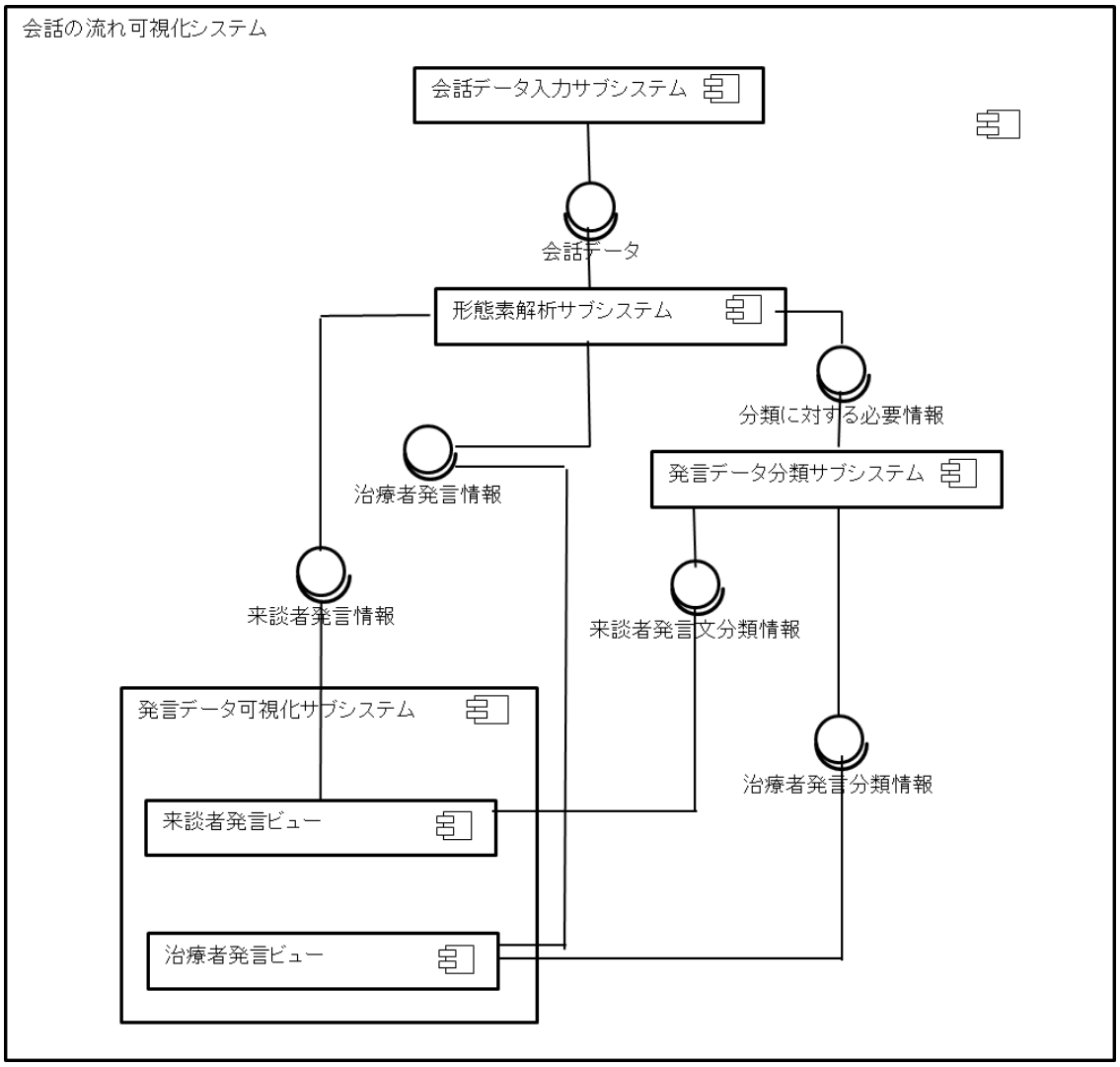

図 4.2: 提案システムの構成

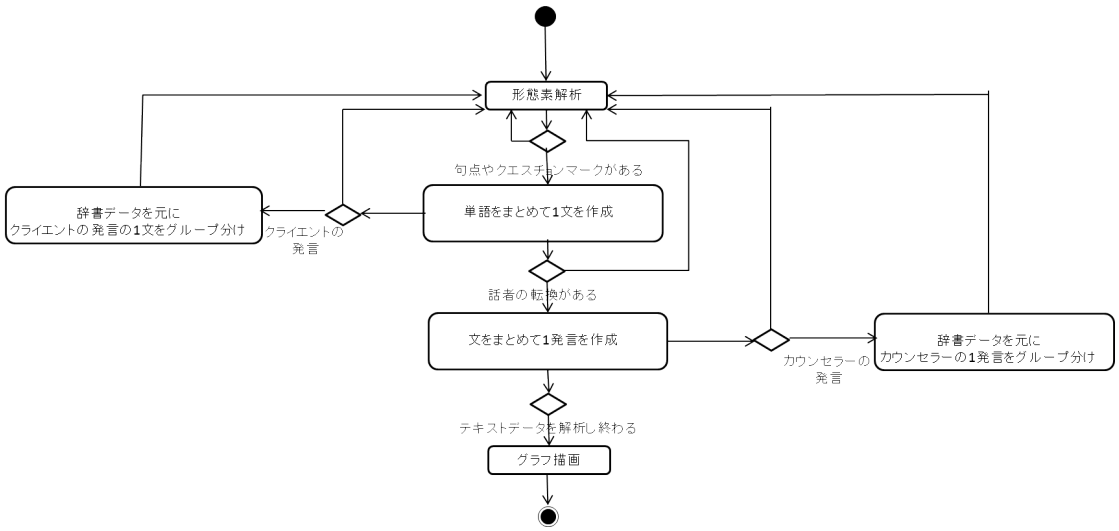

図 4.3: テキストデータ処理のアクティビティ図

図 4.4: 会話データ読み込み前のシステムスクリーンショット

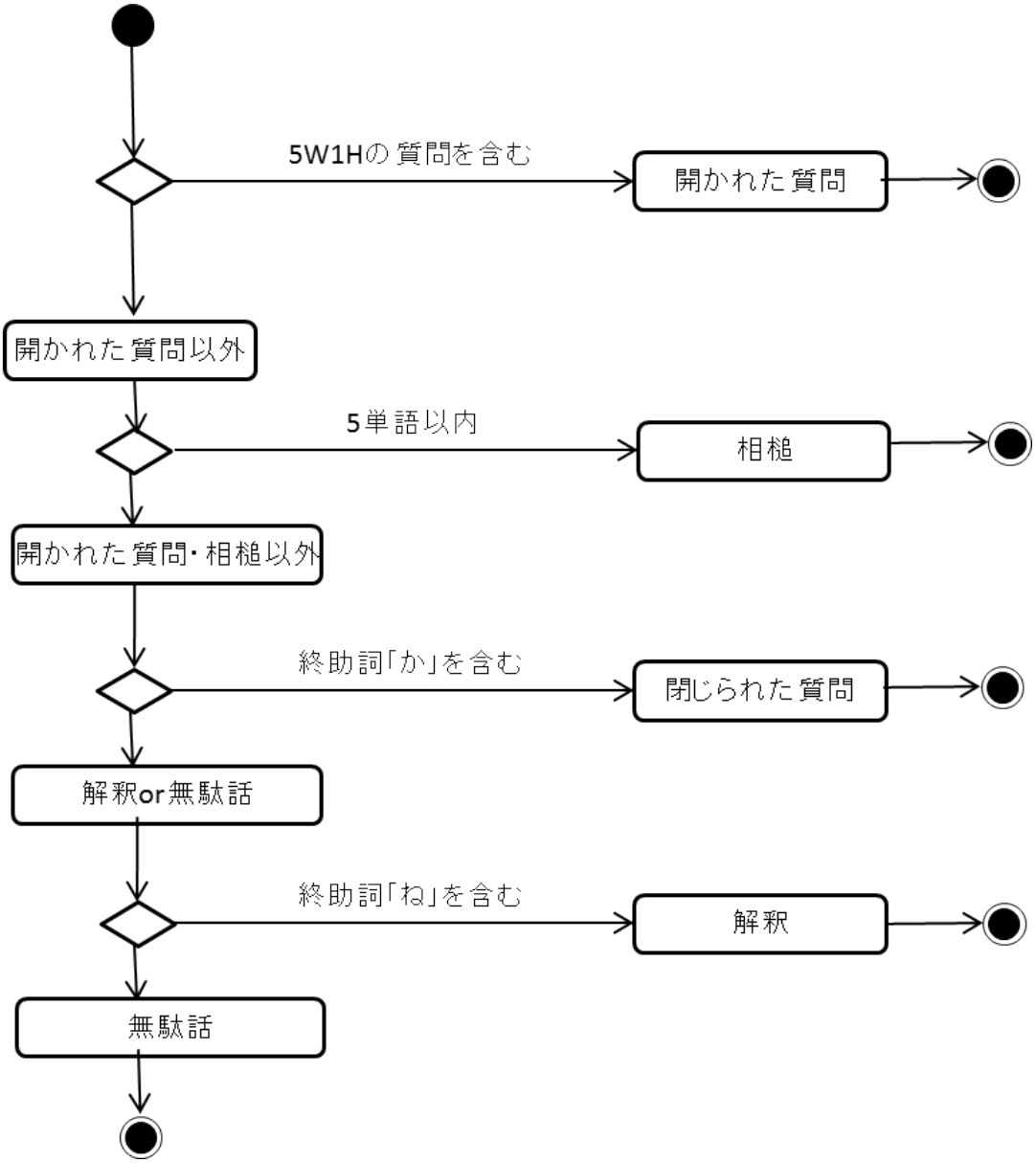

図 4.5: 治療者発言初期分類方法