平 成 30 年 度　卒 業 研 究 論 文

題目　実用的な模擬面接システムのための質疑応答における論理破綻検出に関する研究

指導教員　　服部 峻

提出者　　室蘭工業大学 情報電子工学系学科

氏　名　　清水　康平

学籍番号　　15024083

提出年月日　　平成 31 年 2 月 13 日

目次

図目次

表目次

第 1 章

まえがき

就職活動において，志望者が求める人材であるかを評価するために，多くの企業が面接を行っている．大抵の企業では，志望者を細かく評価するために，面接は複数回にわたって行われている．そのため，就職活動での選考において，面接は他の選考要素に比べても，特に重要な試験であると考えられる．面接において，面接官は学歴や基礎能力，マッチング等に重みを置き，様々な観点によって志望者の評価を行っている．企業によって，面接の形式や面接を実施する面接官に違いがあるが，近年，どの企業でも共通して重視される要素として，基礎能力の一部である論理的思考力が挙げられる．論理的思考力とは，質問の意図をしっかり理解して，筋道を立てて分かりやすく説明ができる能力のことである．質問に対して，回答が的外れであったり，段々と回答が脱線してしまうと論理的思考力が不足していると言える．面接官は複数の質問から，志望者が論理的思考力を持って回答できているかを見抜き，志望者を評価する．論理的思考力は，面接官が志望者についてよく知るために，志望者自身の経験を根拠とする回答が求められていることが多い．そのため，志望者は論理的思考力を持って回答するために，事前に面接に対する準備をしっかりと行う必要がある．しかし，準備を行う上で，志望者が自身の回答を客観的に評価することは難しいため，志望者は回答を改善するのが難しいという問題がある．また，実際の面接を経験しても，面接がどう評価されたのかフィードバックを貰える機会がないことや，従来の模擬面接システムや AI 面接では，回答の分析が行えず，回答の良し悪しを判断することができないため，志望者は自身の回答に問題があったのか判断が難しい．

そこで，回答文に着目して分析を行い，論理破綻を検出することで，ユーザに回答の改善を促すシステムを提案する．現在も，就職活動を支援することを目的として，面接での定番の質問を出題し，ユーザの回答を録音することで，ユーザがその回答を聞き直して改善を行う質問集のようなアプリや，姿勢や発声をより良くして自分自身の印象を良く見せられるようにする支援システムは存在する．しかし，回答文に着目し分析を行うシステムはなく，回答の内容について触れるシステムは存在しなかった．近年の傾向を考えると，回答を分析することでユーザの論理的思考力を評価するシステムがあれば，面接経験が少ない就活生は模擬面接システムで練習を行うことで，より志望者自身を面接官に伝えられるように改善できると考える．そのため，本システムは，「論理的な回答では，回答内の各文の関係性は高いはずである」という

仮説に基づく論理破綻検出により回答を分析する模擬面接システムとなっている．面接での論理的な回答に見られる構成の特徴に着目した接続詞と助動詞による論理構成の推定と，回答文の各文の名詞を抽出し，TF-IDF による重要語の設定と Word2Vec による重要語間のコサイン類似度により，回答文の各文同士の関係性を推定することで，質問に対する回答の論理破綻を判定する．この判定結果をユーザに提示し，その結果から回答を改善して貰う流れになっている．また，面接用の正解セットを用意し，システムによる判定結果が正解セットとどの程度一致しているか評価実験を行う．

第 2 章

関連研究

面接における回答を分析する関連研究は存在しないため，別の文章形式を対象に文を分析する関連研究について調査を行った．関連研究を回答での論理破綻検出にそのまま踏襲することはできないかを考え，できない場合にはどのような問題が発生しているのか考察する．

2.1　小論文の論理構成の把握

まず，小論文の論理構成の把握 [1] では，接続表現と文末モダリティに重みを置いて小論文の論理構成を把握している．しかし，指示代名詞や接続表現，文末モダリティは小論文と面接の回答では異なる．また，小論文の文章量が 800 字〜1,600 字であるのに対し，面接での回答文では，回答時間は長くて凡そ 1 分前後が良いとされており，文章量にすると 300 字前後であるため，小論文と面接の回答では文章形式に違いがある．そのため，小論文での論理構成把握だけでは，十分に面接の回答の論理破綻を検出することができないので，接続表現と文末モダリティによる理構成把握を参考にしつつ，別の手法も交えて回答を分析する必要がある．そこで，接続詞（接続表現）と文末表現（文末モダリティ）に着目することで論理構成を把握していることを参考に，面接での接続詞と文末表現（助動詞）による論理構成推定を論理破綻検出に組み込むことを検討した．

2.2　短答式記述式試験での採点支援

次に，短答式記述式試験での採点支援 [2] では，3 つの判定処理によって短答式記述試験の採点を行う．しかし，短答式記述試験は明確な正解が必ず存在するが，面接の回答では存在しない．そのため，判定処理の一部である上位語・下位語の概念に基づく共起性による文章の正当性の判別を参考に，後述する「論理的な回答では，回答内の各文の関係性は高いはずである」という仮説の関係性推定に応用できないかを考えた．面接の回答では，明確な正解が存在しないため，質問に対するまったく同じ意味をもつ回答であっても，自分自身を幅広く表現することができる特徴がある．そのため，文中に含まれる名詞が他の品詞に比べても，その文自体の意味を表すのに重要な品詞であることが考えられる．しかし，この場合では，共起性だけでは

幅広く名詞の意味をとることができず，面接の回答向けではなかった．それぞれの名詞間での関係性の有無の調査による，各文の関係性の推定は論理判定に有効であると考え，より面接向けに，Word2Vec を用いた名詞の類似度による推定方法を検討する．

2.3　対話破綻検知チャレンジ

最後に，対話破綻検知チャレンジ [3] では，システムでの学習にアノテータがクラス分類したデータセットを用いて破綻検知を行う．データセットは，システムとユーザによる単純な短文でのやり取りの繰り返しに対して，それぞれアノテータがラベル付けしているものである．しかし，面接の回答文では，対話とは文量や形式が異なる．また，面接では人それぞれで回答の良し悪しを判断する観点や材料が異なり，一般な面接官の志望者への評価基準は明確になっていない．そのため，面接向けシステムではアノテータを用いた学習を用いた論理破綻検出は困難であると考えた．そこで，本研究では，アノテーションに依らない論理破綻検出手法を検討する．

2.4　まとめ

以上より，それぞれの研究で対象としている文章形式が異なるため，関連研究をそのまま踏襲することはできなかった．そのため，関係研究で文を分析する上で，文の中で着目している箇所を参考に，面接での回答に応じた独自の手法を用意する．

第 3 章

提案手法

3.1　質疑応答の論理破綻検出の概要

論理的な回答については，人それぞれで論理判定を行う観点や材料が異なるため，明確に定義されていない．そのため，回答の良し悪しを判断するのには人それぞれで違いがある．そこで，本研究における論理的な回答をここで定義する．「論理的な回答では，回答内の各文の関係性は高いはずである」という仮説を立て，論理的な回答であるために必要な 3 つの観点を以下のように定義した．

（1）論理構成推定

論理的な回答の構成が満たされているかを判定するには，品詞に着目し判断する手法がよく用いられる．面接での論理的な回答では，初めに趣旨が存在し，趣旨への根拠と，結論が最後にあることや，根拠が回答者自身の経験した内容であり，時制が過去形であることが多いという特徴がある．また，接続詞はその性質から，文前後の関係性を推定することが可能である．そのため，助動詞の時制と接続詞に着目し，回答が趣旨・根拠・結論で構成されているか推定することで論理判定を行う．また，文中に名詞が存在しないことや，文章量が少な過ぎて具体性がなければ，単純な回答である．単純な回答であれば，論理構成を推定することができないため，回答は論理的ではないと言える．

（2）適正性判定

質問に対する趣旨の内容が一致していなければ，的外れな回答を行っていると考えられる．質問と趣旨の文中に存在する重要語に着目することで，質問に対する適正性を判定する．互いの文で適正でないと判定されると，論理破綻していると言える．

（3）一貫性判定

回答の中でどの文とも関わりを持たない文（孤立文）が存在すると，一貫性がなく不十分である．また，趣旨と結論同士の内容が一致していないと，回答は不十分である．回答文の重要語に着目することで，文同士の関わり合いを推定して論理破綻

判定を行う．孤立文判定と趣旨と結論の一致判定で，どちらかの判定が満たされていなければ，一貫性判定において回答は論理破綻していると言える．

これら 3 つの観点に基づき，システムで論理判定を行う．3 つ全ての観点を満たした場合に論理的な回答として判定し，それ以外の場合であれば論理破綻した回答であると判定する．また，企業によっては面接での質問に違いがあるため，本研究では，多くの企業で共通して出題されている質問での回答の分析を対象としている．例として，これらの観点で判断した場合での，論理的な回答と論理的ではない回答の各々の回答例を示す．

論理的な回答の例

（例）「入社後にやりたい仕事は？」

『私は、笑顔ナンバーワンと言われるような接客のプロフェッショナルになりたいです。私は幼いころから御社の店舗を利用しており、親切で丁寧、かつ洗練されたプロの接客に強い尊敬と憧れを抱いていました。そこから私も接客業に興味を持ち、高校・大学時代に飲食店でアルバイトをしていました。そこでお客様と接する際に注意することや必要な気遣いなどを学び、人と接することがより楽しくなりました。これらの経験を活かし、御社でプロの販売員になれるよう精進していく所存です。』

この例では，3 つ全ての観点を満たしており，論理的な回答であると言える．

論理的ではない回答の例

（例）「あなたの長所は何ですか？」

（1）論理構成推定が満たされていない

『特にありません。』

『私が学生時代に最も打ち込んだことはボランティア活動です。』

この例では，どちらの文も文章量が少なく，単純な回答となってしまっているため，論理構成推定が満たされていない．

（2）適正性判定が満たされていない

『私は人に役立てることを大切にしたいと考えます。ボランティア活動で不登校の児童・生徒の家庭訪問を行っていました。彼らが学校に通うことができた瞬間は非常に嬉しく、やりがいを感じました。御社でも、常に人に役立てるために自分のできることを探しながら、仕事をしていきたいと考えています。』

この例では，質問文『あなたの長所は何ですか？』と趣旨『私は人に役立てることを大切にしたいと考えます。』との間に関係性がないため，適正性判定が満たされていない．

（3）一貫性判定が満たされていない

『私の長所は協調性があることです。問題を抱えているサークルメンバーのノートや資料作成を手伝ったり、情報や人を紹介したりしていました。問題解決につながったときは嬉しかったです。私は入社後、海外事業に携

わりたいと強く考えております。大学では 1 年間休学し、アメリカに留学しておりました。』

この例では，結論『大学では 1 年間休学し、アメリカに留学しておりました。』が他の文との間に関係性がないため，結論文は孤立文であるので，孤立文判定を満たさない．また，趣旨『私は人に役立てることを大切にしたいと考えます。』と結論『大学では 1 年間休学し、アメリカに留学しておりました。』との間に関係性がないため，趣旨と結論が一致していない．そのため，一貫性判定を満たしていない．

3.2　システムの処理

本研究での模擬面接システムの処理を図 3.1 に示す．

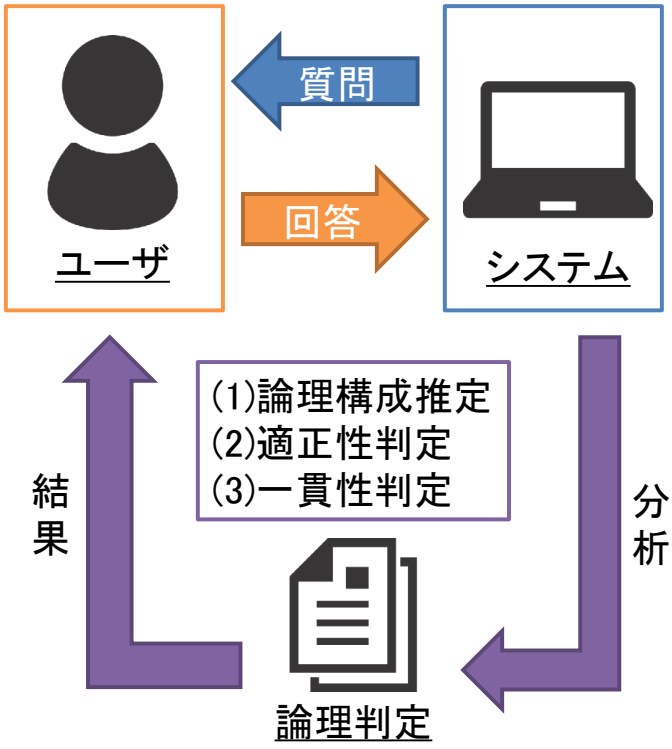

図 3.1　システム全体の流れ

システムは質問集からユーザに対して質問をランダムに出題する．今回は，様々な企業で共通して出題されている 10 件の質問を用意している．ユーザは，質問に対する回答を行い，その回答をシステムで論理破綻していないか分析し，論理判定結果を出力する．この工程を用意している質問の数だけ繰り返す流れとなっている．

また，今回では実装には至らなかったが，最終目標としているシステムのインターフェースを図 3.2 に示す．

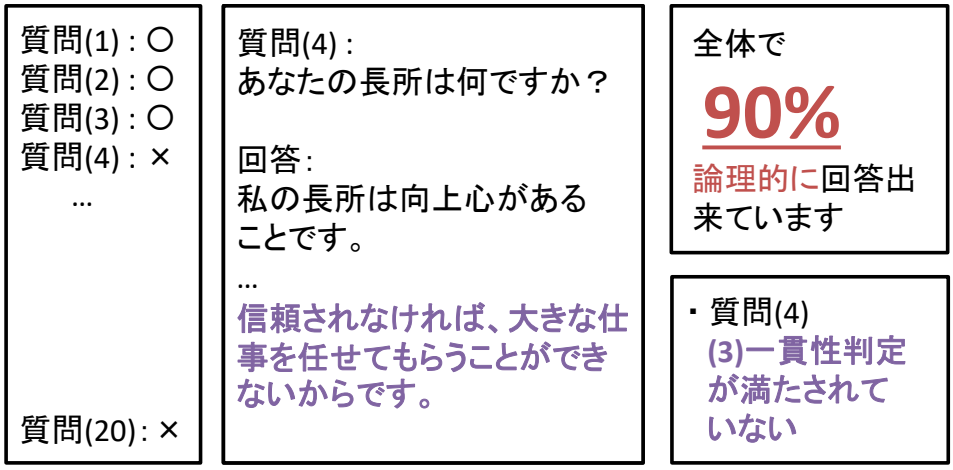

図 3.2　システムのインターフェース

模擬面接終了後に，回答全体でどの程度論理的に答えられていたかの結果をユーザに提示する．論理破綻していた回答には，破綻原因をユーザが見返せるよう提示し，その質問に対する論理的な回答の正解例を提示することで，ユーザは回答の改善を行うように考えている．

3.3　形態素解析エンジン MeCab の活用

文中での名詞の抽出と後述するデータセットと学習モデルの作成のために，形態素解析エンジンである MeCab を活用する．また，MeCab のシステム辞書として，mecab-ipadic-NEologd[4] を用いる．これは，MeCab 用のシステム辞書であり，Web 上から得た新語に対応しているため，固有名詞表現に強い特徴がある．MeCab により抽出された名詞を対象に，TF-IDF の重要度に基づく重要語の設定や，Word2Vec での重要語間の類似度を求める．与えられた文章に対して，実際に形態素解析を行い，各文から名詞を抽出した例を図 3.3 に示す．

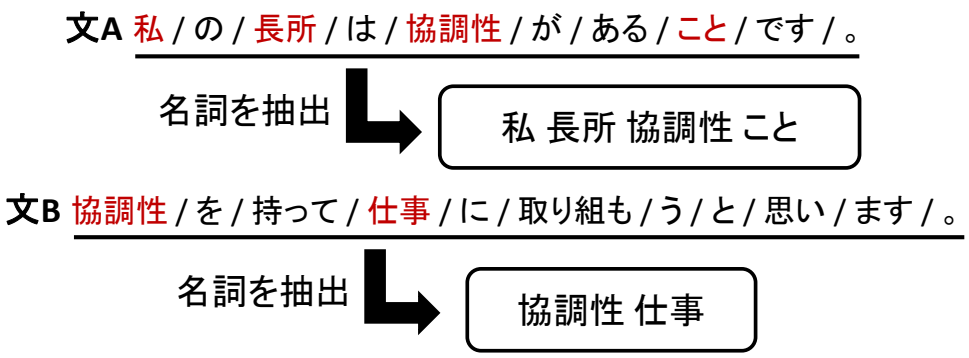

図 3.3　形態素解析による名詞の抽出の例

3.4　TF-IDF による重要語の抽出

TF-IDF は，複数の文書がある場合に，文書内でのある単語の出現頻度とある単語が含まれる文書の割合の逆数により，文中に含まれる各単語の重要度を求められる手法である．面接の質問・回答をまとめたデータセットを用意し，回答内の各文に含まれる名詞の重要度を求める．各文から設定された個数分の重要度の高い名詞を抽出し，抽出された名詞を重要語とする．重要語を検出する流れを図 3.4 に示す．重要語に着目することで，見落としがなく文同士の関係性を判断できると考える．後述する評価実験の結果により，重要語の最良な個数について考察する．

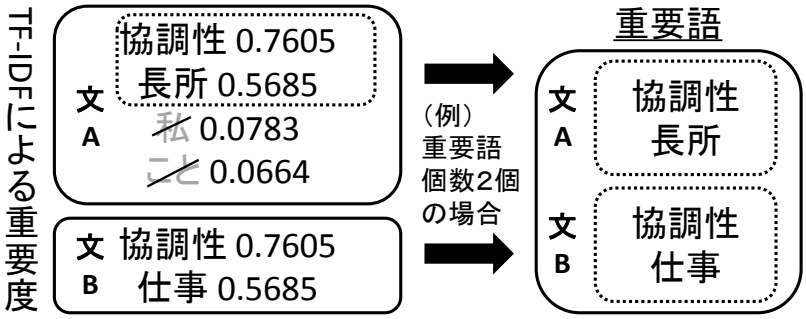

図 3.4　TF-IDF による重要語の抽出

この図 3.4 の例では，重要語の個数は 2 個と設定されている．そのため，文 A での名詞の中で重要度の高い上位 2 件の『協調性』,『長所』が文 A の重要語として抽出される．また，文 B ではそもそも名詞が 2 個しかないため,『協調性』,『仕事』が文 B の重要語として抽出される．

3.5　Word2Vec による文同士の関係性推定

Word2Vec は，大量のテキストデータを解析し，各単語の意味をベクトル表現可能な手法である [5]．学習モデルを用いることにより，単語間の類似度を求めることができる．TF-IDF により重要語を抽出し，重要語間で類似度を求める．基準値を設定し，各文から抽出された重要語間の類似度が基準値以上の場合に，文同士の関係性があると判定する．重要語間での組み合わせは複数あるが，関係性のある組み合わせが 1 つもなければ，文同士の関係性がないと判定する．重要語による関係性推定結果 R の次の数式より算出する．文 A と文 B での重要語の個数を N，対象とする文 A の重要語 w_i $(i = 1, 2, ..., N)$ と文 B の重要語 w_j $(j = 1, 2, ..., N)$ の類似度を $S(w_i, w_j)$ とし，設定した基準値を θ とする．

$$R = \sum_{i=1}^{N} \sum_{j=1}^{N} F(w_i,\ w_j)$$

$$F(w_i,\ w_j) = \begin{cases} 1 & \text{if } S(w_i,\ w_j) \geq \theta, \\ 0 & \text{otherwise.} \end{cases}$$

$S(w_i,\ w_j) \geq \theta$ となる組み合わせ $(w_i,\ w_j)$ が 1 組もなければ，文 A と文 B との間に関係性がないと判定し，それ以外では，文 A と文 B との間に関係性があると判定される．

また，Word2Vec により，文同士の関係性を推定する流れを図 3.5 に示す．後述する評価実験の結果により，基準値の最良な値について考察する．

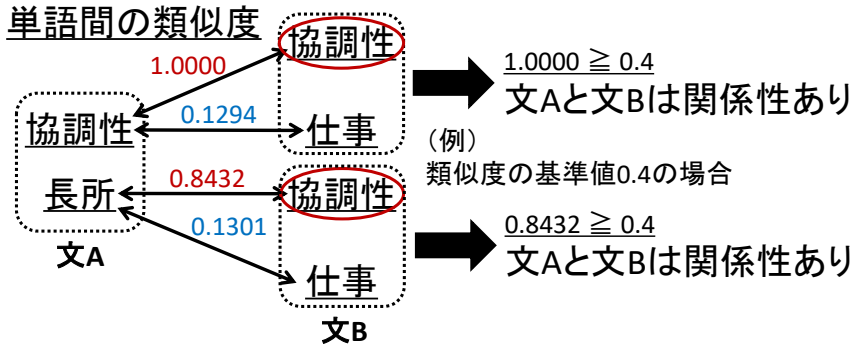

図 3.5　重要語間の類似度による文同士の関係性推定

この図 3.5 の例では，類似度の基準値 θ は 0.4 と設定されている．文 A の重要語『協調性』と文 B の重要語『協調性』で類似度を求めると，1.0000 であり基準値 0.4 以上であるため，文 A と文 B との間に関係性があると判定される．また，文 A の重要語『長所』と文 B の重要語『協調性』で類似度を求めると，0.8432 であり基準値 0.4 以上であるため，文 A と文 B との間に関係性があると判定される．もし，文 A に重要語『協調性』が存在せず，重要語『長所』のみとした場合では，基準値 0.4 以上の重要語の組み合わせは存在しないため，文 A と文 B との間に関係性がないと判断される．

3.6　面接データセットと学習モデルの作成

TF-IDF と Word2Vec では，それぞれ計算にデータセットが必要となる．そのため，就活サイト「就活会議」[6] から実際の企業で出題された面接での質問・回答を収集し，収集したデータを用いて，TF-IDF での面接データセットと Word2Vec での学習モデルを作成する．面接データセットと学習モデルの作成の流れを図 3.6 に示す．「就活会議」に投稿されている企業 2,287 社を対象に，28,270 件の質問・回答が収集できた．収集した質問・回答に対して，MeCab を用いて名詞のみを抽出し，TF-IDF での面接データセットと Word2Vec 用の学習モデルを作成した．

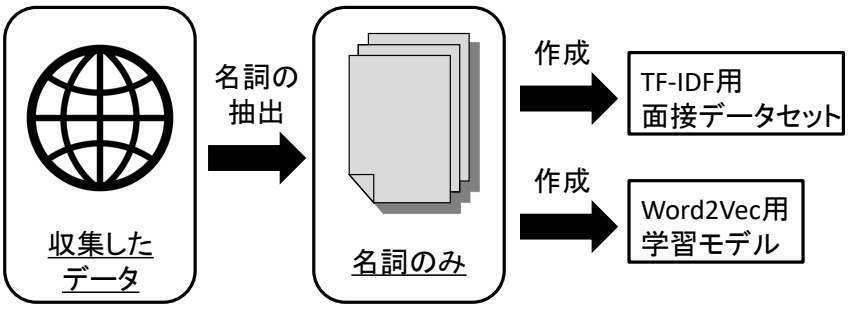

図 3.6　面接データセットと学習モデル作成の流れ

3.7　質疑応答の論理破綻検出の実装

3.1 節で述べた，回答が論理的であるための 3 つの観点毎に，論理的か否かを判定する手法を以下のように実装する．

（1）論理構成推定

論理構成推定で対象としている接続詞の例と助動詞の時制による分類を以下に示す．

接続詞の例（全 23 種）

「だが」 「しかし」 「また」 「そのため」 「つまり」

助動詞

時制が現在形 「です」 「ます」 「ません」

時制が過去形 「でした」 「ました」

文頭に接続表現がある場合と，助動詞の時制が前の文と一致する場合には，前の文と関係性があると推定でき，各文毎の関係を判断して回答全体を趣旨・根拠・結論という 3 つの構成に分ける．回答での最初の文を趣旨とし，次の文から根拠と結論の推定を行う．前の文と関係性がない文が検出されれば，次の構成に切り替わったと判断する．その中で，回答に不足している構成があれば，回答は論理的でないと判定する．回答の各文における助動詞の時制の変化による構成の切り替え判断の例を図 3.7 に示す．

また，単純な回答の判定も行う．回答文から何か一つでも名詞を抽出できているか，趣旨の文字数は 10 文字以上であるか，回答文は 3 文以上であるかを判断する．もし，これらの条件が満たされていなければ，論理的な回答ではないと判定する．

（2）適正性判定

質問文と，（1）論理構成推定で回答を構成に分けた趣旨の部分とに含まれる重要語を設定された個数分それぞれ抽出し，互いの類似度を求める．質問文と趣旨で抽出された重要語間で類似度を求め，全ての組み合わせで類似度が設定した基準値より

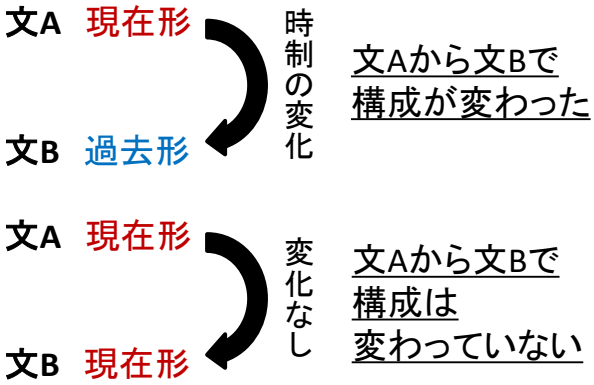

図 3.7　助動詞の時制による構成の変化の例

低ければ，回答は論理的でないと判定する．適正性判定での重要語の個数と基準値について，後述する評価実験にて，最良な値を設定する．

（3）一貫性判定

回答の全ての文中にある重要語を抽出し，各文の重要語間の類似度を求める．注目した文と他の文との類似度が設定した基準値より低ければ，その文は孤立文であると判定する（孤立文判定）．同様に，趣旨と結論の類似度が基準値より低ければ，趣旨と結論が一致していないと判定する（趣旨と結論の一致判定）．この例を図 3.8 に示す．回答中で最初の文（1 文目）が趣旨であり，最後の文（5 文目）が結論である．この回答中で，4 文目は他の全ての文との間に関係性がないため，4 文目は孤立文である．また，趣旨（1 文目）と結論（5 文目）との間に関係性がないため，趣旨と結論は一致していない．従って，この例では論理的ではない回答であると判定される．一貫性判定での孤立文判定と趣旨と結論の一致判定で，それぞれ用いる重要語の個数と基準値を，後述する評価実験にて，最良な値を設定する．

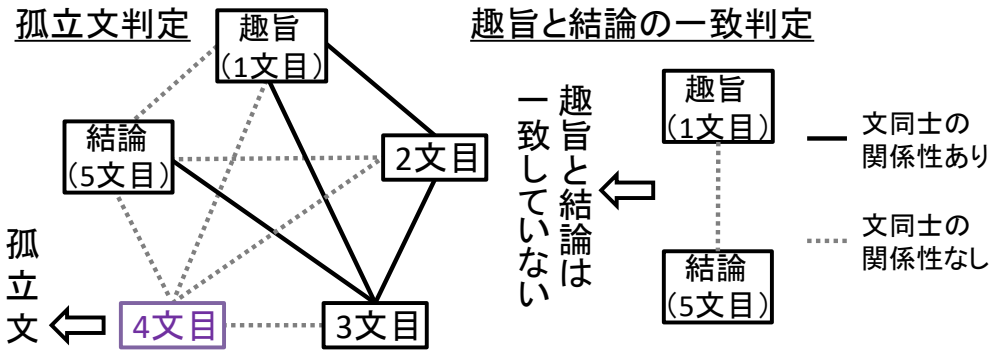

図 3.8　孤立文を含むかつ趣旨と結論が一致していない例

また，以上の 3 つの観点に基づく処理から成る，本システムでの論理破綻検出のアルゴリズムを図 3.9 に示す．

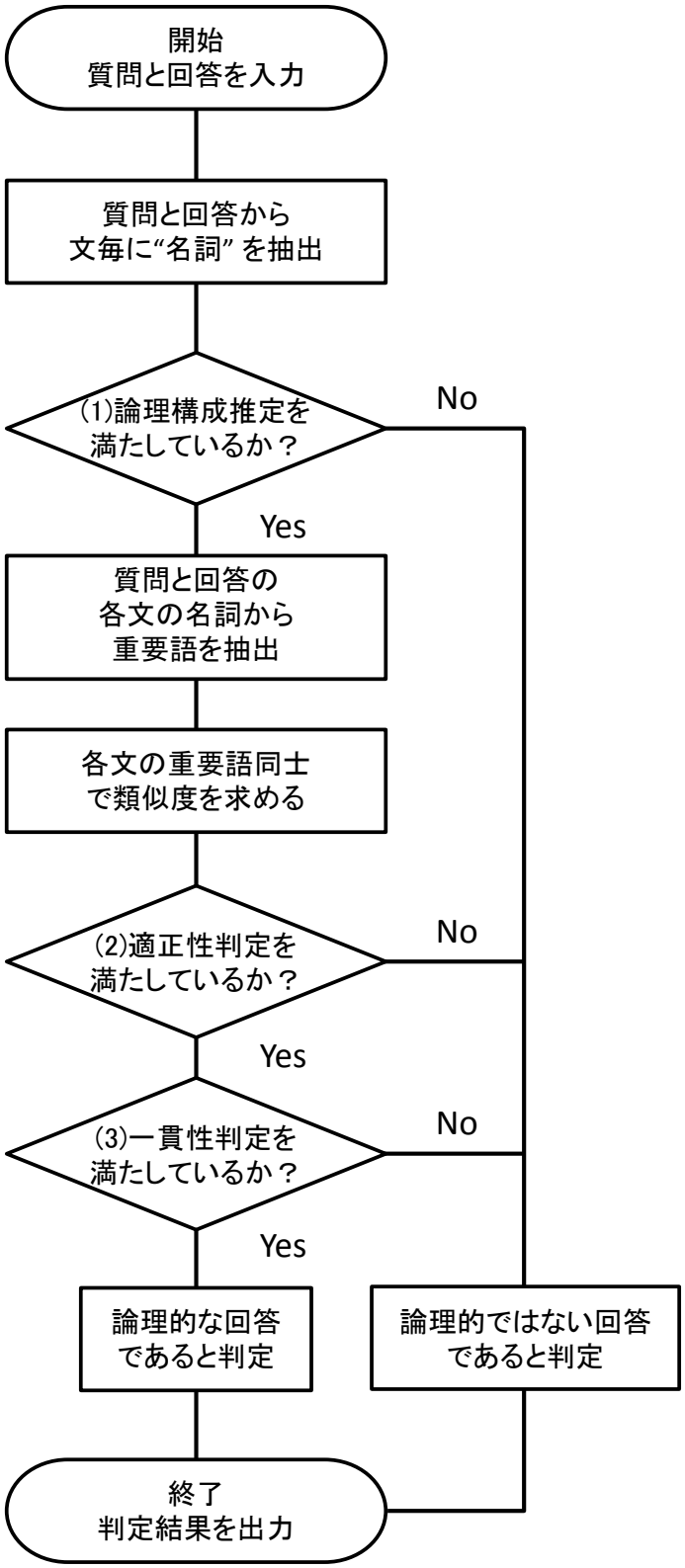

図 3.9　論理破綻検出のアルゴリズム

第 4 章

システム評価と今後の課題

4.1　実験概要

まず，（2）適正性判定と（3）一貫性判定において，重要語の個数 N と，重要語間の類似度に関する基準値 θ という 2 種類のパラメータの最良な組み合わせについて調査した．特に，一貫性判定では，孤立文判定と，趣旨と結論の一致判定のそれぞれで最良の F 値となるパラメータの組み合わせを調査した．また，これら 3 つの最良なパラメータの組み合わせをシステムで用いて，システムでの判定結果が正解セットとどの程度一致しているか評価実験を行う．これらの実験結果より，本システムの精度と今後の課題について考察した．

4.2　正解セットの作成

作成した正解セットの構成を図 4.1 に示す．

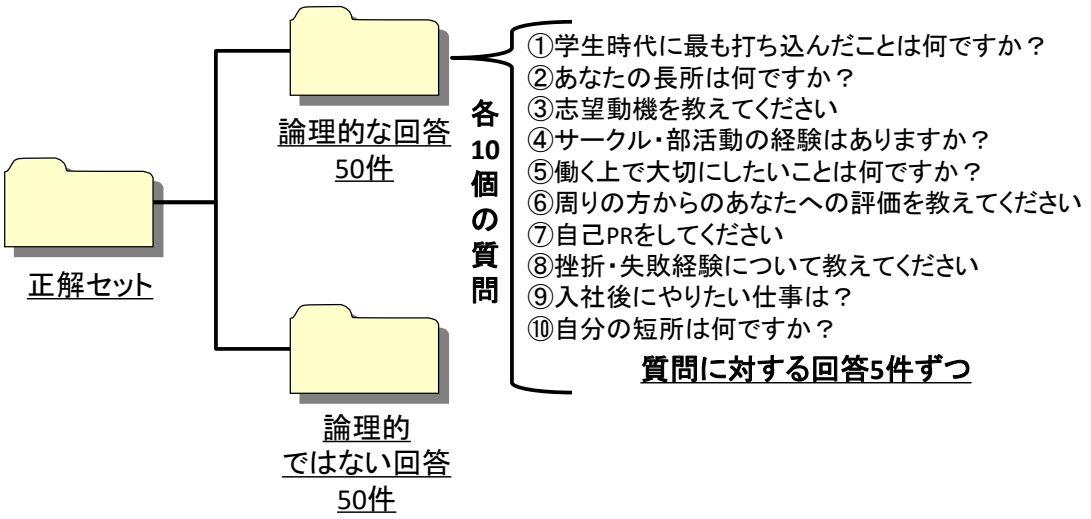

図 4.1　正解セットの構成

論理的な回答 50 件と論理的ではない回答 50 件の計 100 件の回答を用意した．これらの回答は検証用として，ネット上の回答例を参考に筆者が作成したものである．多くの企業で共通

して出題されている質問 10 件に対して，論理的な回答では 5 件ずつ用意し，論理的ではない回答では，(1) 論理構成推定を満たさない回答 2 件，(2) 適正性判定を満たさない回答 1 件，(3) 一貫性判定における孤立文が存在する回答 1 件，趣旨と結論が異なる回答 1 件の計 5 件ずつ用意した．

4.3　最良な重要語の個数と基準値

(2) 適正性判定と，(3) 一貫性判定における孤立文判定，及び趣旨と結論の一致判定において，最良な重要語の個数 N と，類似度に関する基準値 θ について最良のパラメータを調査した．作成した面接用データセットでの 1 文に含まれる名詞の数は平均で約 8 個はあった．文中に重要語が多いと Word2Vec での比較対象が増えてしまい，膨大な検証時間がかかる．そのため，本研究の評価実験では，重要語の個数は最大 8 個までとし．重要語 1〜8 個，基準値 0.1〜0.9 でそれぞれ正解セットとの F 値を求めた．F 値はシステムと正解セットで論理的であると判定された回答に着目し，適合率，再現率を用いて計算する．適合率，再現率，F 値の計算方法を以下に示す．それぞれの判定での F 値を表 4.1〜表 4.3 に示す．また，それぞれの判定での 3 次元グラフを図 4.2〜図 4.4 に示す．

$$\text{適合率} := \frac{\text{システムと正解セットの両方で論理的である回答}}{\text{システム判定で論理的である回答全体}}$$

$$\text{再現率} := \frac{\text{システムと正解セットの両方で論理的である回答}}{\text{正解セットで論理的である回答全体}}$$

$$\text{F 値} := \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}}$$

表 4.1　（2）適正性判定での重要語の個数 N と基準値 θ に依る F 値の変化

θ \ N	1 個	2 個	3 個	4 個	5 個	6 個	7 個	8 個
0.1	0.774	0.895	0.907	0.917	0.917	0.917	0.917	0.917
0.2	0.729	0.848	0.865	0.897	0.897	0.907	0.917	0.917
0.3	0.700	0.787	0.848	0.893	0.914	0.925	0.935	0.935
0.4	0.632	0.753	0.863	0.909	0.920	0.941	**0.962**	**0.962**
0.5	0.563	0.716	0.809	0.848	0.851	0.875	0.887	0.898
0.6	0.522	0.649	0.791	0.831	0.831	0.844	0.844	0.835
0.7	0.413	0.571	0.701	0.765	0.765	0.780	0.780	0.771
0.8	0.387	0.551	0.630	0.667	0.667	0.684	0.684	0.675
0.9	0.333	0.507	0.571	0.611	0.611	0.630	0.630	0.622

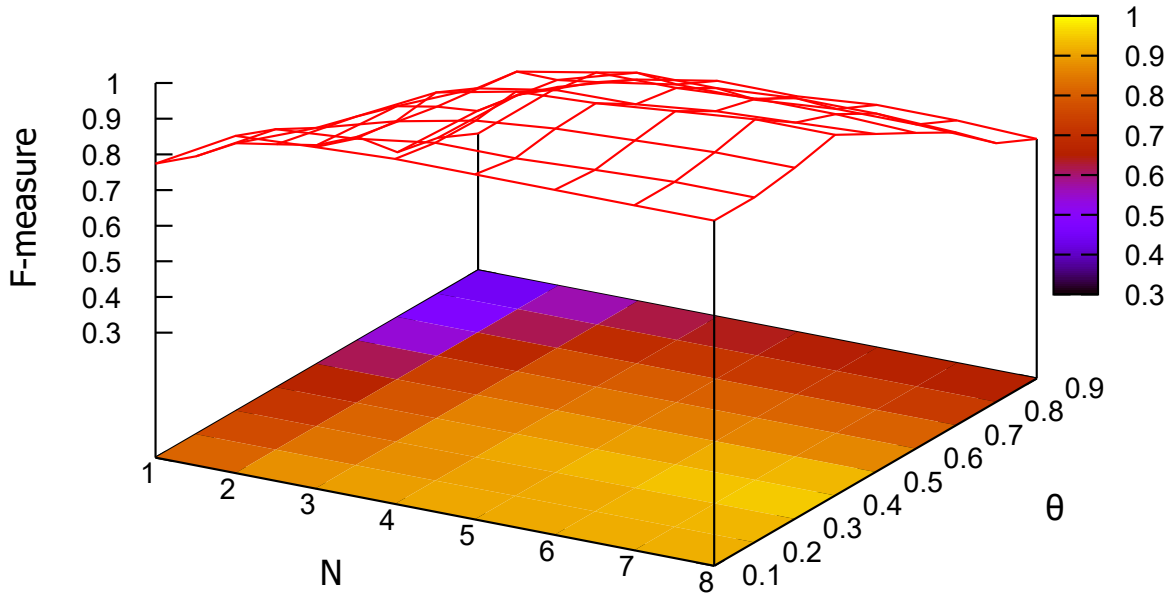

図 4.2　（2）適正性判定での重要語の個数 N と基準値 θ と F 値に依る 3 次元グラフ

表 4.2　（3）一貫性判定における孤立文判定での重要語の個数 N と基準値 θ に依る F 値の変化

θ \ N	1 個	2 個	3 個	4 個	5 個	6 個	7 個	8 個
0.1	0.907	0.917	0.909	0.909	0.909	0.909	0.909	0.909
0.2	0.916	0.917	0.909	0.909	0.909	0.909	0.909	0.909
0.3	0.906	0.917	0.909	0.909	0.909	0.909	0.909	0.909
0.4	0.871	**0.926**	0.917	0.909	0.909	0.909	0.909	0.909
0.5	0.866	0.916	**0.926**	0.917	0.917	0.917	0.917	0.917
0.6	0.866	0.916	**0.926**	**0.926**	**0.926**	**0.926**	**0.926**	**0.926**
0.7	0.817	0.874	0.895	0.895	0.895	0.895	0.906	0.906
0.8	0.750	0.863	0.885	0.885	0.895	0.895	0.895	0.906
0.9	0.736	0.871	0.882	0.882	0.885	0.885	0.885	0.895

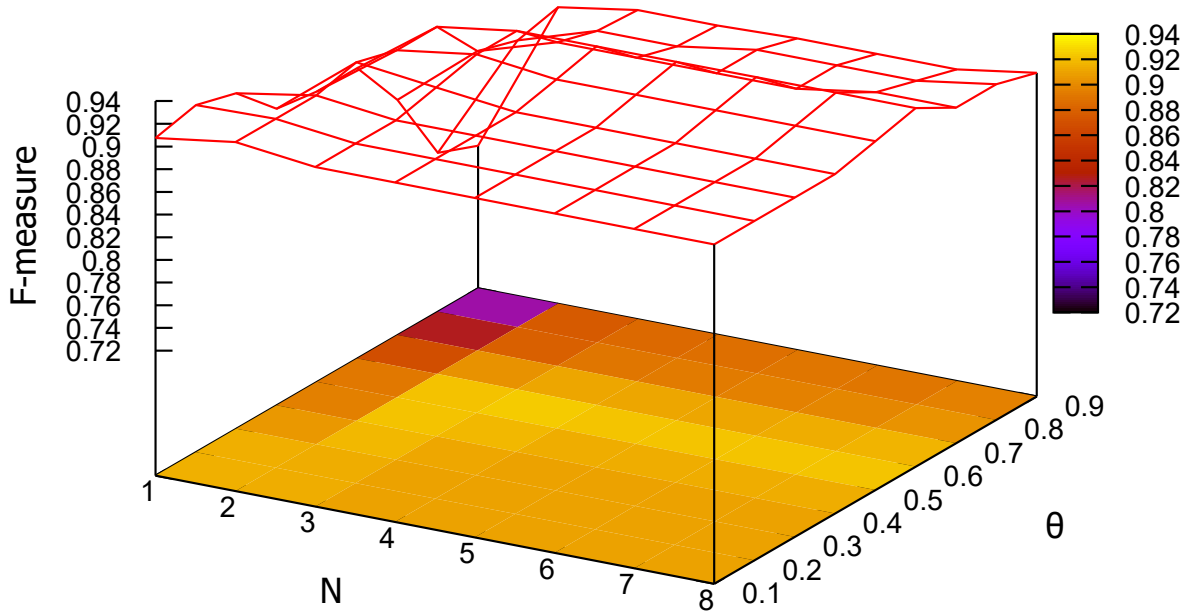

図 4.3　（3）一貫性判定における孤立文判定での重要語の個数 N と基準値 θ と F 値に依る 3 次元グラフ

表 4.3　（3）一貫性判定における趣旨と結論の一致判定での重要語の個数 N と基準値 θ に依る F 値の変化

θ \ N	1 個	2 個	3 個	4 個	5 個	6 個	7 個	8 個
0.1	0.851	**0.917**	0.909	0.909	0.909	0.909	0.909	0.909
0.2	0.833	**0.917**	0.909	0.909	0.909	0.909	0.909	0.909
0.3	0.706	0.885	0.889	0.899	0.899	0.899	0.909	0.909
0.4	0.729	0.837	0.854	0.857	0.868	0.879	0.879	0.879
0.5	0.683	0.800	0.843	0.835	0.835	0.835	0.846	0.857
0.6	0.641	0.711	0.800	0.804	0.804	0.816	0.828	0.828
0.7	0.575	0.667	0.714	0.721	0.698	0.727	0.742	0.747
0.8	0.471	0.533	0.642	0.667	0.674	0.705	0.719	0.725
0.9	0.471	0.507	0.568	0.615	0.625	0.659	0.675	0.659

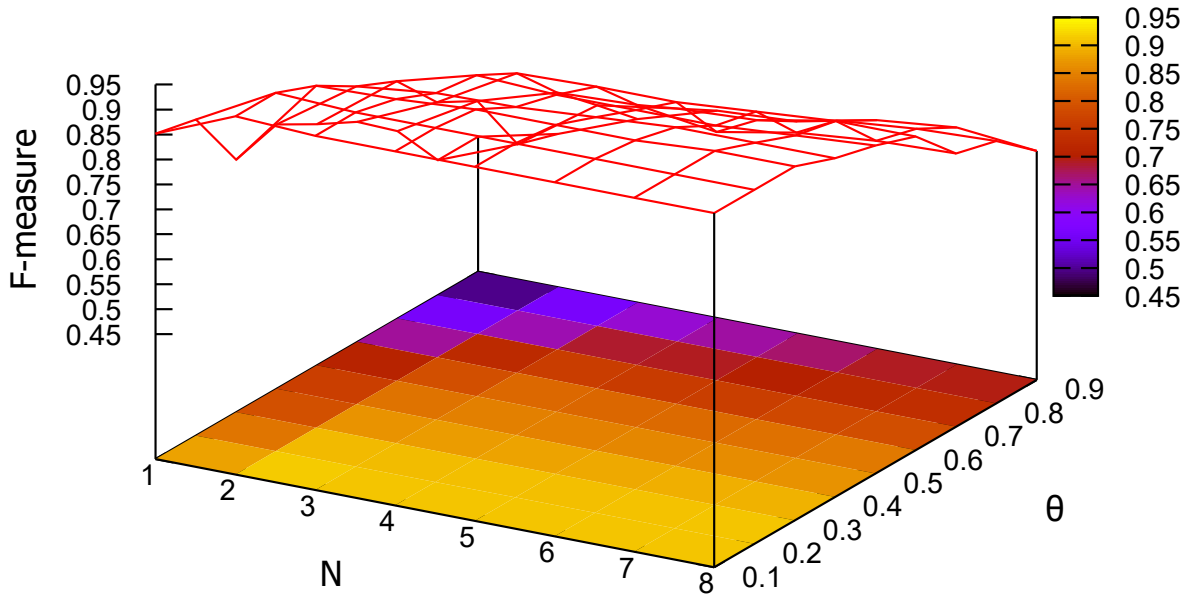

図 4.4　（3）一貫性判定における趣旨と結論の一致判定での重要語の個数 N と基準値 θ と F 値に依る 3 次元グラフ

まず，表 4.1 を見ると，（2）適正性判定で最も高い F 値は 0.962 であり，重要語 7 個または 8 個で，基準値 0.4 の場合であった．どの判定であっても，重要語の個数に着目すると，ある個数からは F 値は安定する傾向となっている．そのため，適正性判定では，重要語 8 個，基準値 0.4 を最良なパラメータの組み合わせとする．同様に，表 4.2 を見ると，（3）一貫性判定における孤立文判定で最も高い F 値は 0.926 であり，複数の場合でこの値が出ている．その中で，基準値 0.6 で最も高い F 値が多く存在していた．そのため，孤立文判定では，値が安定している重要語 6 個，基準値 0.6 を最良なパラメータの組み合わせとする．また，表 4.3 を見ると，（3）一貫性判定における趣旨と結論の一致判定で最も高い F 値は 0.917 であり，重要語 2 個で，基準値 0.1 または基準値 0.2 の場合であった．趣旨と結論の一致判定での基準値では，基準値が低くなれば F 値は高くなる傾向がある．そのため，趣旨と結論の一致判定では，重要語 2 個，基準値 0.1 を最良なパラメータの組み合わせとする．

4.4　パラメータを最適化されたシステムの論理破綻判定結果

4.3 節で設定した最良なパラメータを用いて，各判定とシステム全体において正解セットに対する適合率，再現率，F 値を求めた．各判定におけるそれぞれの値を，表 4.4〜表 4.8 に示す．

表 4.4　（1）論理構成推定での判定結果

正解＼システム	論理的である	論理的ではない
論理的である	42	8
論理的ではない	0	20

適合率	1.000
再現率	0.840
F 値	0.913

表 4.5　（2）適正性判定での判定結果

正解＼システム	論理的である	論理的ではない
論理的である	50	0
論理的ではない	4	6

適合率	0.926
再現率	1.000
F 値	0.962

表 4.6　（3）一貫性判定での判定結果

正解＼システム	論理的である	論理的ではない
論理的である	50	0
論理的ではない	16	4

適合率	0.758
再現率	1.000
F 値	0.862

表 4.7　（3）における孤立文判定での判定結果

正解＼システム	論理的である	論理的ではない
論理的である	50	0
論理的ではない	8	2

適合率	0.862
再現率	1.000
F 値	0.926

表 4.8　（3）における趣旨と結論の一致判定での判定結果

正解＼システム	論理的である	論理的ではない
論理的である	50	0
論理的ではない	9	1

適合率	0.847
再現率	1.000
F 値	0.917

表 4.9　システム全体での判定結果

正解＼システム	論理的である	論理的ではない
論理的である	42	8
論理的ではない	13	37

適合率	0.764
再現率	0.840
F 値	0.800

どの判定であっても，ある程度良好な F 値が得られた．3 つの観点の中で，（2）適正性判定が最も F 値が高く，（3）一貫性判定が最も F 値が低い結果となった．また，（1）論理構成推定では，正解セットの論理的ではない回答を全て正しく判定できていたが，論理的な回答を若干誤って論理的ではない回答と判定してしまっていた．他 2 つの判定では，正解セットの論理的な回答を全て正しく判定できていたが，論理的ではない回答を若干誤って論理的な回答と判定してしまっていた．それぞれの判定とシステムでの判定結果に対して考察する．

4.5　実験を通じての考察

評価実験で得られた結果から，各判定処理とシステム全体での精度について詳しく考察する．(1) 論理構成推定では，正解セットの論理的ではない回答 20 件を問題なく判定することができていた．しかし，論理的な回答 50 件に対しては，論理的ではない回答であると 8 件が誤った判定が行われていた．8 件の回答について調べてみると，いずれの回答も時制にまったく変化がないことが発覚した．他の判定処理では，論理的な回答については正しく判定が可能であったため，論理構成推定のみが論理的な回答への分析で問題が発生していることになる．時制の変化がないため，推定処理自体は行えるが，論理構成の要件を満たさず，論理構成推定が正しく行われていないと推測する．時制の変化がない回答例を以下に示す．

時制のない回答の例

（例）「周りの方からのあなたへの評価を教えてください」

> 『私は周りから「慎重な人」だとよく言われます。自分なりにしっかりとプランを練った上で、どうすれば効率良く行動できるかを考えております。私は周りの雰囲気に流されず、冷静に物事を処理します。そのため、周りから意見を求められたり、相談に乗ってくれと依頼されることも多いです。』

この回答の例では，全ての文の助動詞において，『ます』『です』しか存在せず，回答内に時制が現在形の文しか存在しない．そのため，時制の変化がないために論理構成推定が正しく判定処理を行うことができていなかった．また，この質問に対する他の回答例でも，ほとんどの回答の時制に変化が見られなかった．質問によっては，回答に時制に変化が見られないケースがあることが判明した．論理構成推定では，時制による論理構成の推定が単純な構成であったため，回答の時制のパターンに応じた細かいルールを追加することで，この問題を解消できると考える．他にも，回答文の文章量に制限を設けたり，他の品詞にも着目する等の複雑な構成にすることで，より精度を向上させることが可能であると考える．

次に，(2) 適正性判定では，正解セットの論理的な回答 50 件を問題なく判定することができていた．しかし，論理的ではない回答に 10 件に対しては，論理的な回答であると 4 件が誤った判定が行われていた．また，(3) 一貫性判定では，正解セットの論理的な回答 50 件を問題なく判定することができていたが，同じく，論理的ではない回答に 20 件に対しては，論理的な回答であると 16 件が誤った判定が行われていた．そのため，文同士の関係性に基づく判定処理単体では，論理的ではない回答に対してあまり精度良く判定できていないなかった．誤って判定が行われていた回答について調べてみると，本来は関係性のない文同士に対して，重要語間の類似度が基準値以上になっていた．これは，文同士での重要語が幅広い意味を持つ名詞であった場合に，簡単に重要語間の類似度が基準値以上となってしまうため，誤った判断が行われることが原因と推定できる．適正性判定において，誤った判定がされた回答を例に，重要語間で類似度が基準値 0.4 以上の組み合わせを以下に示す．

（2）適正性判定において誤った判定がされた例

質問『サークル・部活動の経験はありますか？』

回答『私は高校受験に失敗し、第二志望の学校に進学しました。』

重要語『部活動』と『高校受験』の類似度：0.6087

重要語『部活動』と『失敗』の類似度：0.4481

重要語『部活動』と『学校』の類似度：0.4225

重要語『経験』と『失敗』の類似度：0.4143

この例は，本来は論理的ではない回答であるため，重要語間の類似度が基準値 0.4 以上となる組み合わせは存在してはいけない．しかし，重要語間の類似度が基準値 0.4 以上の組み合わせが 4 つも存在するため，論理的な回答であると判定が行われてしまう．重要語間で基準値を満たす組み合わせが 1 つでも存在すれば，文同士の関係性があると判定する仕様となっているため，検出する重要語の個数に応じて，さらに細かいルールを追加することで，この問題を解消できると考えられる．他にも，Word2Vec 用学習モデルでの学習対象をさらに増やすことで，幅広い意味を持つ名詞に対しても，重要語間でより正確な類似度を算出することが可能であると考える．

最後に，システム全体では，論理的な回答 50 件に対して 42 件を正しく判定でき，論理的ではない回答 50 件に対して 37 件を正しく判定できていた．そのため，システム全体ではある程度正しく判定処理を行うことができており，それぞれの判定処理を併用することで，システム全体でより良い精度が得られることが判明した．

これらの問題に対して，個別に対応していくことにより，システムによる判定結果の精度がより改善できると考えられる．

第 5 章

むすび

本研究では，品詞と文の関係性に基づく質疑応答の論理破綻検出について提案した．論理的な回答 50 件と論理的ではない回答 50 件の計 100 件の回答で作成された正解セットにおける，(2) 適正性判定と (3) 一貫性判定における孤立文判定，及び趣旨と結論の一致判定での重要語の個数 N と，重要語間の類似度に関する基準値 θ の最良なパラメータの組み合わせについて調査し，これらの最良なパラメータを用いたシステムに採用した．また，3 つの観点によるシステム全体の判定として，F 値で 0.800 の精度が得られた．そのため，「論理的な回答では，回答内の各文の関係性は高いはずである」という仮説に基づく，重要語と品詞に着目した (1) 論理構成推定，(2) 適正性判定，(3) 一貫性判定の 3 つの判定処理による論理破綻検出は有効であり，システムでもある程度の精度で人間と同様に論理判定が可能であると言える．また，本研究は人の発言を対象に論理破綻検出を行うため，面接での回答に対する論理判定をそのまま踏襲することはできないが，本研究を応用することにより，会議やコールセンターでの対話における論理破綻検出等にも活用することができると考える．

今後は，各観点に対して，新たに細かいルールを追加し，回答の様々なパターンに対応できるようにすることで，より精度を向上させる．本研究の実験では，正解セットの件数が少なく，最良な F 値が複数個存在していたので，今後は，正解セットの件数を増やし，再度評価実験を行うことで，よりシステムの精度を上げられると考える．また，本システムをユーザに体験して貰えていないため，システムに対するユーザの評価を調査する必要もある．ユーザが実際の面接に近い模擬面接が体験できるように，システムに音声認識による回答文の入力を考えている．この音声認識には，Google の音声認識 API である Google Cloud Speech[7] を用いる．ここで，音声認識の精度について問題があるが，面接の回答では，面接官にはっきりと内容を伝える必要があるため，日常の会話や雑談等の用途で使う場合よりも比較的精度良く使えると考える．しかし，面接を想定して作られた API ではないため，ユーザの発言の言い直しや言い間違えを検知することはできない．この問題に対して，API の精度を検証しつつ，必要に応じて対策を講じる必要がある．音声認識を実装し，実際に本システムでユーザに模擬面接を経験して貰い，面接の内容や論理破綻結果へのユーザの満足度を調査する必要があると考える．

謝辞

本研究に際して，様々なご指導を頂きました服部峻助教に厚く御礼申し上げます．また，日常の議論を通じて多くの知識や示唆を頂いた服部研究室の皆様にも深く感謝の意を表します．

参考文献

[1] 石岡 恒憲, “日本語小論文の論理構成の把握とその図式表現,” 人工知能学会論文誌, vol.23, no.5, pp.303–309 (2008).

[2] 石岡 恒憲, “コンピュータ上で実施する記述式試験—エッセイタイプ，短答式，マルチメディア利用について—,” 電子情報通信学会誌, vol.99, no.10, pp.1005–1011 (2016).

[3] 東中 竜一郎, “対話破綻検出チャレンジ 2　The Dialogue Breakdown Detection Challenge 2” 人工知能学会 言語・音声理解と対話処理研究会, vol.78, pp.64–69 (2016).

[4] mecab-ipadic-NEologd : Neologism dictionary for MeCab, https://github.com/neologd/mecab-ipadic-neologd/blob/master/README.ja.md (2019).

[5] Word2Vec : 発明した本人も驚く単語ベクトルの驚異的な力 - DeepAge, https://deepage.net/bigdata/machine_learning/2016/09/02/word2vec_power_of_word_vector.html (2019).

[6] 【就活会議】新卒採用/インターン/面接の評判がわかる口コミサイト, https://syukatsu-kaigi.jp (2019).

[7] Cloud Speech-to-Text - 音声認識 — Cloud Speech-to-Text API — Google Cloud, https://cloud.google.com/speech-to-text/?hl=ja (2019).