

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

清水 康平[†] 荒澤 孔明[†] 渡邊 稜平[†] 服部 峻^{††}

^{†,††}室蘭工業大学 ウェブ知能時空間研究室 〒050-8585 北海道室蘭市水元町 27-1

E-mail: [†]{15024083,18096001,18043050}@mmm.muroran-it.ac.jp, ^{††}hattori@csse.muroran-it.ac.jp

あらまし 就職活動において面接は、企業が求める人材を獲得するための重要な試験である。面接では、様々な観点から志願者を精査しているが、特に重視される要素に論理的思考力がある。近年では、志願者が論理的思考力を持って回答が来ているかを見抜く質問が多く出題されている。しかし、実際の面接ではフィードバックを貰える機会は無いため、論理的な回答が来ていると評価されたのかは回答者自身では判断しにくい。また、既存の模擬面接システムでは、回答の論理破綻を判断出来ないため、志願者は回答を改善することが難しいという問題がある。そこで、本稿では、自身では判断しにくい回答の論理破綻を自動的に検出して、ユーザの回答改善を支援することを目的とし、回答内の重要語と、単語の品詞（特に接続詞と副助詞）に着目し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、より実用的な模擬面接システムを提案する。

キーワード 模擬面接, 質疑応答, 機械学習, Word2Vec, TF-IDF

Logical Failure Detection in Question-and-Answer Sessions Towards A Practical Mock Interview System

Kohei SHIMIZU[†], Komei ARASAWA[†], Ryohei WATANABE[†], and Shun HATTORI^{††}

^{†,††} Web Intelligence Time-Space (WITS) Laboratory, Muroran Institute of Technology
27-1 Mizumoto-cho, Muroran, Hokkaido, 050-8585, Japan

E-mail: [†]{15024083,18096001,18043050}@mmm.muroran-it.ac.jp, ^{††}hattori@csse.muroran-it.ac.jp

Abstract In job hunting, oral interviews are important examinations for job hunters to appeal to their targeted companies and for companies to acquire human resources who they want to seek. In oral interviews, interviewers of companies scrutinize each job hunter from various viewpoints. Among them, “logical thinking” is one of especially-important factors. In recent years, the questions to find out whether or not each job hunter can answer them with logical thinking are often asked in oral interviews. However, it is difficult for a job hunter to judge whether or not s/he was evaluated to be able to answer logically by her/his targeted companies, because there is no opportunity for her/him to receive feedback about it in actual interviews. Meanwhile, it is difficult for a job hunter to improve her/his answers using the existing mock interview systems, because they cannot detect a logical failure in her/his answers. Therefore, this paper proposes a more practical mock interview system that aims at helping job hunters to improve their answers by themselves, and that automatically detects logical failures in a job hunter’s answer to a company’s question by estimating the relationship between sentences focusing on the important words and the part of speech (in particular, conjunction and adverbial particle) of words in her/his answer.

Key words Mock Interview, Question and Answering, Machine Learning, Word2Vec, TF-IDF

1. はじめに

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。面接では、様々な観点で志望者を精査しているが、近年では、特に重視される要素として論理的思考力が挙げられる。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来ているかを見抜き、志望者を評価する。論理的思考力は、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く、志望者は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、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は無く、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、そ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無く、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、回答を分析し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。

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」という仮説に基づいている。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構成になっている。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。

2. 提案手法

2.1 質疑応答の論理破綻検出の概要

小論文を対象に論理的か評価を行うシステム [1] は存在しているが、質疑応答に関する文章に対しては未だに不十分である。小論文と質疑応答では、文の長さや形式も大きく異なっているため、小論文を対象とした既存のシステムをそのまま踏襲することはできない。そのため、従来の小論文の論理構成把握とは異なる手法で質疑応答の論理破綻を検出する必要がある。

質疑応答の論理判断では、人それぞれによって、回答を判断する観点には違いがあり、その定義については様々な意見がある。そのため、本稿における、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、文の数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、一貫性が無く不十分である。また、趣旨と結論同士の内容が一致していないとすると、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。

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断し、それ以外の場合であれば論理破綻した回答であると判断する。

例として、これらの観点を判断した場合での、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)「学生時代に打ち込んだことはなんですか？」

『居酒屋で接客のアルバイトに打ち込みました。元気だけは誰にも負けないように、人一倍大きな声で挨拶をすることを心がけて働きました。このお店に来れば楽しく元気になる、とお客様に思っていた良かったです。3年目にはバイトのリーダーを任せられ、お店の売上にも貢献することができました。』

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。』

(2) 適正性判定が満たされていない

『私は人に役立つ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・生徒の家庭訪問に行っていました。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、常に人に役立つ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。』

(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。』

2.2 システムの処理

本稿での模擬面接システムの処理を図1に示す。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、質問集に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する。この工程を用意

している質問の数だけ繰り返す。面接終了後、システムは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。また、論理破綻していた回答については、破綻原因をユーザが見返せるよう提示する流れとなっている。

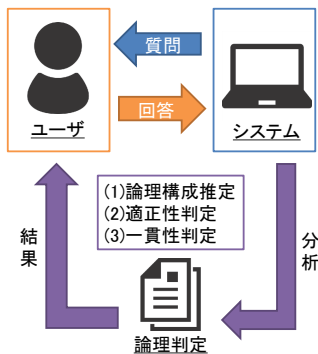


図1 システム全体の流れ

2.3 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 [2] 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 での重要語同士の類似度を求める。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2に示す。

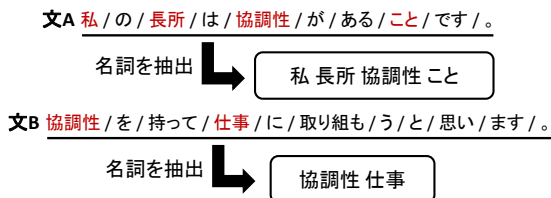


図2 形態素解析による名詞の抽出の例

2.4 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、単語の出現頻度と逆文書頻度により、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に示す。この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。

2.5 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 [3]。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、重要語同士で類似度を求める。基準値

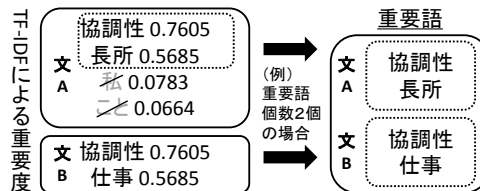


図3 TF-IDF による重要語の抽出

を設定し、各文から抽出された重要語同士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。Word2Vec により、文同士の関係性を推定する流れを図4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。

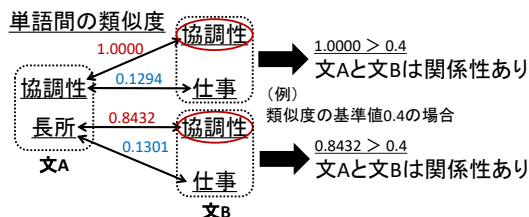


図4 重要語同士の類似度による文同士の関係性推定

2.6 面接データセットと学習モデルの作成

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。本稿では、就活サイト「就活会議」[4] 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、このデータを用いて、面接データセットと学習モデルを作成した。収集できた企業2,287社を対象に、28,270件の質問・回答を得ることができた。面接データセットと学習モデルの作成の流れを図5に示す。取得した質問・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。

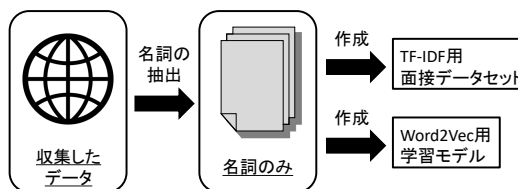


図5 面接データセットと学習モデル作成の流れ

2.7 質疑応答の論理破綻検出の実装

2.1 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 「ました」「でした」

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。前の文と関係性が無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断する。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6に示す。

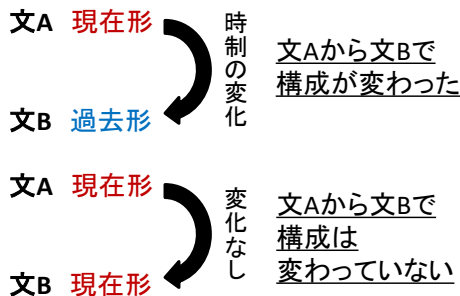


図6 助動詞の時制による構成の変化の例

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定する。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断する。

(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同士で類似度を求め、全ての組み合わせで類似度が設定した基準値より低ければ、回答は適正性が無いと判定する。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、最適な値を設定する。

(3) 一貫性判定

回答の全ての文中にある重要語を抽出し、各文の重要語同士の類似度を求める。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7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関係性が無いため、4文目は孤立文である。また、趣旨(1文目)と結論(5文目)で文同士の関係性が無いため、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、最適な値を設定する。

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8に示す。

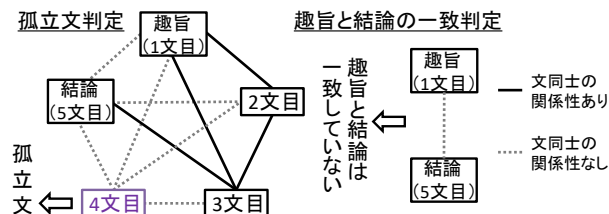


図7 孤立文を含むかつ趣旨と結論が一致していない例

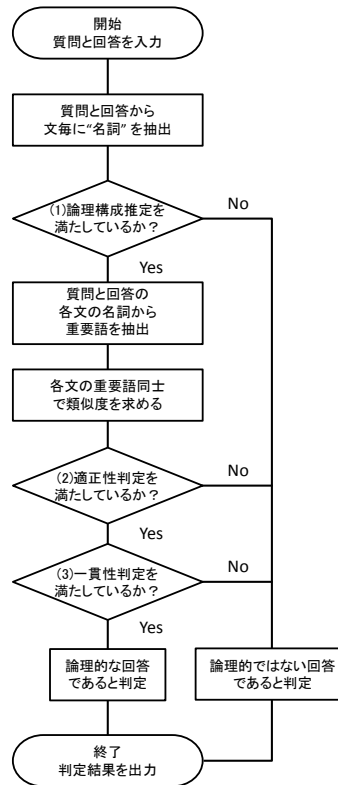


図8 論理破綻検出のアルゴリズム

3. 評価実験

3.1 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適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、最適なパラメータの組み合わせによる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。

3.2 正解セットの作成

作成した正解セットの構成を図9に示す。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。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。

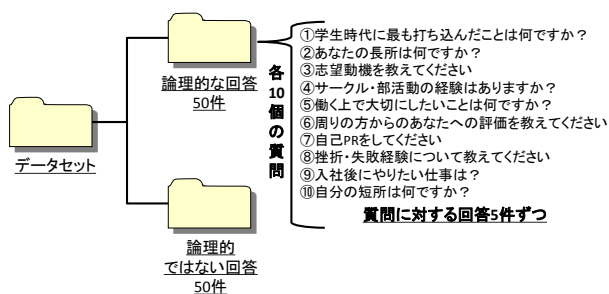


図9 正解セットの構成

3.3 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, 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 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, 本稿の評価実験では, 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値を図10~図12に示す.

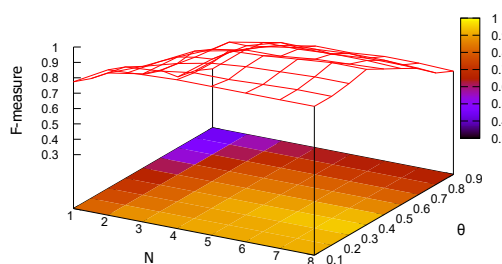


図10 (2) 適正性判定での重要語の個数 N と基準値 θ に依る F 値の変化

| $\theta \backslash N$ | 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 |

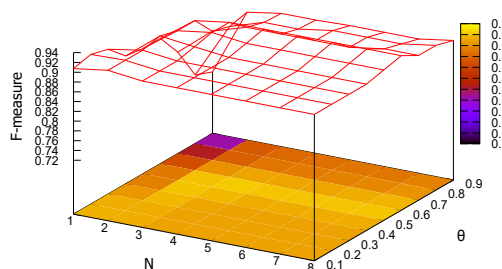


図11 (3) 一貫性判定における孤立文判定での重要語の個数 N と基準値 θ に依る F 値の変化

| $\theta \backslash N$ | 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 |

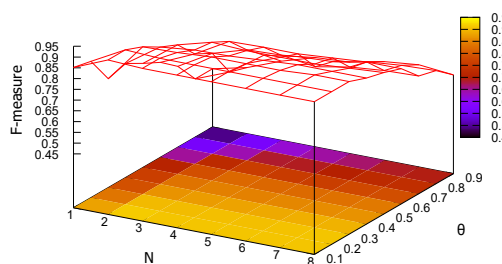


図12 (3) 一貫性判定における趣旨と結論の一致判定での重要語の個数 N と基準値 θ に依る F 値の変化

| $\theta \backslash N$ | 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 |

表 1 (1) 論理構成推定での判定結果

| システム | | 論理的である | | 論理的ではない | |
|------|---------|--------|-------|---------|-------|
| | | 適合率 | 1.000 | 再現率 | 0.840 |
| 正解 | 論理的である | 42 | 8 | F 値 | 0.913 |
| | 論理的ではない | 0 | 20 | | |

表 3 (3) における孤立文判定での判定結果

| システム | | 論理的である | | 論理的ではない | |
|------|---------|--------|-------|---------|-------|
| | | 適合率 | 0.862 | 再現率 | 1.000 |
| 正解 | 論理的である | 50 | 0 | F 値 | 0.926 |
| | 論理的ではない | 8 | 2 | | |

表 5 (3) 一貫性判定での判定結果

| システム | | 論理的である | | 論理的ではない | |
|------|---------|--------|-------|---------|-------|
| | | 適合率 | 0.758 | 再現率 | 1.000 |
| 正解 | 論理的である | 50 | 0 | F 値 | 0.862 |
| | 論理的ではない | 16 | 4 | | |

表 2 (2) 適正性判定での判定結果

| システム | | 論理的である | | 論理的ではない | |
|------|---------|--------|-------|---------|-------|
| | | 適合率 | 0.926 | 再現率 | 1.000 |
| 正解 | 論理的である | 50 | 0 | F 値 | 0.962 |
| | 論理的ではない | 4 | 6 | | |

表 4 (3) における趣旨と結論の一致判定での判定結果

| システム | | 論理的である | | 論理的ではない | |
|------|---------|--------|-------|---------|-------|
| | | 適合率 | 0.847 | 再現率 | 1.000 |
| 正解 | 論理的である | 50 | 0 | F 値 | 0.917 |
| | 論理的ではない | 9 | 1 | | |

表 6 システム全体での判定結果

| システム | | 論理的である | | 論理的ではない | |
|------|---------|--------|-------|---------|-------|
| | | 適合率 | 0.764 | 再現率 | 0.840 |
| 正解 | 論理的である | 42 | 8 | F 値 | 0.800 |
| | 論理的ではない | 13 | 37 | | |

$$\begin{aligned} \text{適合率} &:= \frac{\text{システムと正解セットの両方で論理的である回答}}{\text{システム判定で論理的である回答全体}} \\ \text{再現率} &:= \frac{\text{システムと正解セットの両方で論理的である回答}}{\text{正解セットで論理的である回答全体}} \\ \text{F 値} &:= \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \end{aligned}$$

まず、図 10 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、図 11 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、図 12 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。

3.4 最良な値によるシステムによる論理破綻判定結果

3.3 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 1~表 6 に示す。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。

3.5 実験を通じての考察

以上の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、論理的ではない回答を問題なく判定することができていたが、論理的な回答では若干誤った判定がされることがあった。これは、回答の各文が同じ時制であっても、論理的な回答となるケースが存在したことで誤った判定が行われていたことが原因であると推測できる。時制による論理構成の変化の推定が単純であったため、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。

(2) 適正性判定と (3) 一貫性判定では、論理的な回答に対しては、問題なく判定することができていたが、論理的な回答ではない回答では若干誤った判定が行われることがあった。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、文同士で関係性が無くても、類似度が基準値を超えてしまい、誤った判定がされてしまうことが原因であると推定できる。基準値を満たす重要語の組み合わせが 1 つでもあれば、文同士の関係性があると判断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。

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。

4. むすびに

本稿では、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。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、品詞と文の関係性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。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。本稿の実験では、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。また、より実用的な模擬面接システムのため、システムを音声認識を対応させて、ユーザの発言から回答文を音声で入力できる機能の追加を考えている。

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] mecab-ipadic-NEologd: Neologism dictionary for MeCab, <https://github.com/neologd/mecab-ipadic-neologd/blob/master/README.ja.md> (2018).
- [3] 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 (2018).
- [4] 【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> (2018).