ユーザの精神疾患度合いを自動的に診断する対話システムの構築

高木 優　　増田 優　　皆川 正浩
大内 啓樹　秋間 大輔　近藤 雅芳
松本 裕治
奈良先端科学技術大学院大学　情報科学研究科

{takagi.yu.tn0, masuda.yu.mj0, minakawa.masahiro.mb2,ouchi.hiroki.nt6,
akima.daisuke.zu0, kondo.masayoshi.jy5, matsu}@is.naist.jp

1　はじめに

近年日本において，精神疾患により医療機関にかかっている患者数は継続的に増加しており，平成 11 年は 204 万人であったのが平成 23 年は 320 万人となっている [1]．いくつかの精神疾患は，発症後に適切な治療を受ければ症状を緩和出来る一方で，その治療効果は発症から治療開始までの未治療期間の長さが影響する．日本では，精神疾患への無知や精神科受診に対する偏見を原因とした，未治療期間の長さが問題視されており，日本政府も精神疾患啓蒙を目的とした事業を各地で実施している．

自然言語処理の応用研究において，現実の医療課題をテーマとした研究が近年発表されている [2]．一方で，精神疾患の早期診断を目的とした研究は著者らの知る限り存在しない．本研究では，精神疾患の早期診断を目的に，ユーザとの自然な対話から精神疾患の重症度を診断する対話システムを構築する．本対話システムは，臨床の現場でも用いられている自己記入式の重症度判定シートを元に，ユーザとの対話から得られた情報を用いて重症度を推定する．評価実験の結果，本対話システムによるユーザの重症度推定結果は，同一被験者による自己記入式の重症度判定シートの結果を高い精度で近似出来ることが確認された．この結果は，システムとの自然な対話から精神疾患の診断を適切に行える可能性を示唆するものであり，医療現場における診断補助や遠隔医療などといった分野での応用が期待される．

2　関連研究

文書から著者の性別や年齢などを推定する研究は，以前から行われている [3][1]．これらの研究は文書から著者の属性を推定する点で本研究と共通するが，推定対象が性格や年齢という一般的な指標である点が本研究とは異なる．精神疾患との関連では，患者特有の文章表現の異常を同定する研究や [5]， 自由作文からうつ病の重症度を推定する研究 [4] などが存在する．これらの研究は精神疾患患者の文書を扱う点で本研究と共通している一方，患者が一方的に記述した文書のみを対象としており，双方向的な対話により取得した文書を利用していない点が本研究とは異なる．

3　対話システム概要

本研究で構築するシステムの概要を図 1 に示す．システムからの質問文の発話，ユーザからの応答文の取得，取得した応答文の意味理解，意味理解を元にしたユーザの重症度評価値の計算，以上の結果をもとにした相槌および質問文を含んだ発話文の生成，を 1 サイクルとし，ユーザの重症度が確定するまでこのサイクルを回す．本システムでは，各質問に対して複数の選択肢から回答する自己記入式の重症度判定シートを元に，質問文を人手で作成している．そのため，そのようなシートが存在する疾患であれば本システムによる診断が可能である．今回の研究では，対象とする精神疾患を，日本で最も患者数が多いうつ病とした．使用する自己記入式の重症度判定シートとしては，厚生労働省が Web 上で公開している簡易抑うつ症状尺度（Quick Inventory of Depressive Symptomatology: QIDS）を用いた [2]．QIDS は 16 個の質問からなり，各質問に 4 択で回答する．その後，定められた計算方法によって回答者の重症度が計算される．

[1] http://www.mhlw.go.jp/bunya/shougaihoken/kokoro/dl/02.pdf
[2] http://www.mhlw.go.jp/bunya/shougaihoken/kokoro/dl/02.pdf

表 1: 本研究で使用した発話文と実際の応答文の例

発話形式	発話例	応答例
感情極性型	眠っていても落ち着かなかったり、眠りが浅くて早く起きてしまったことはありますか？	よくあります
	自分が他の人に迷惑をかけていると強く思いますか？	全くそんなことはない
数値評価型	寝るまでに何分ぐらいかかってますか？	20 分ぐらいです
	昼寝も含めると 1 日平均で何時間ぐらい寝ていますか？	11 時間は寝ているかな
相槌	眠れないとき、私は桜の木を千本数えますよ。	-

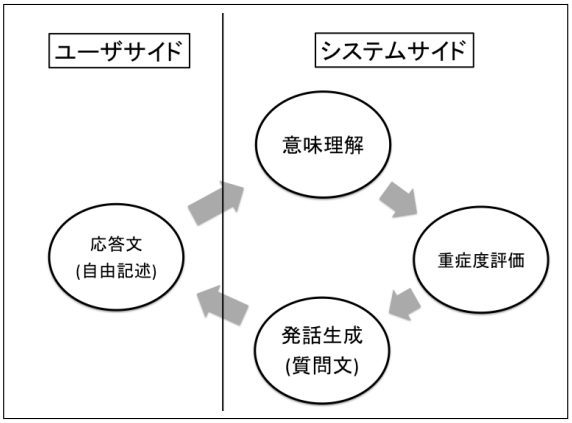

図 1: 本研究で開発したシステムの構造

3.1 入力文の意味解析方法

意味理解モジュールでは，形態素解析システム Juman Ver.7.0[3] と構文・格解析システム KNP Ver.4.1[4] を用いる．Juman で文を単語分割し品詞情報を同定した後，KNP で単語間の係り受け関係や各述語の項の同定を行う．各質問に対して期待される応答には，数値情報を答える応答と肯定・否定を答える応答の二種類がある．例えば，数値を問う質問として「昼寝も含めると 1 日平均で何時間ぐらい寝ていますか.」などがあり，これに対する応答文から具体的な数値情報を抽出する．肯定・否定を問う質問として「周りの人や，いろんな活動について，普段と比べて興味が薄れていると感じますか.」があるが，この質問に対する応答が肯定・否定のどちらを意味しているかを識別する必要がある．本システムでは応答文の肯定・否定を識別するために，各単語の持つ感情極性を考える．感情極性値として -1 から $+1$ の実数値を各単語に割り振っている単語感情極性対応表 [5] を使用し [7]，文内の各単語が持つ感情極性値の合計が正の値であればユーザの意図は肯定的，負の値であればユーザの意図は否定的であると解釈する．具体的な評価方法としては，応答文中の各単語と単語感情極性対応表に記載されている単語のマッチングを行い，マッチすればその感情極性値を抽出する．システムは各応答文における感情極性値の総和を算出する．その際，“ない” や “ぬ” といった否定を表す語が含まれていれば極性を反転させる．最終的な結果が正の値であれば，その応答文を肯定的と認識し，負の値であれば否定的と認識する．なお，単語感情極性対応表については，予備実験により確認された応答文の特徴を踏まえ，以下の基準で一部改編した．

1. 質問文内に含まれている名詞および動詞は，評価に影響する情報を含まないことが多いため，それらが感情極性辞書に含まれていた場合はその極性値を 0 へと変更した．

2. 工藤 [6] が挙げている思考・知覚動詞は，評価に影響する情報を含まないことが多いため，単語感情極性対応表に含まれていた単語の極性値を 0.000001 に変更した．

上記の 2. に関して説明を加えると，例えば “〜と思いますか” などの肯定・否定を問う質問に対して “思う” などが単独で応答として用いられる場合は肯定的と認識するのが望ましい．逆に，“〜をどう思いますか” などの WH 型の質問では，“思う” の極性値は 0 になるのが望ましい．例えば “最悪だと思う” という文において，“思う” の極性値が正で “最悪” の極性値が負であり，“思う” の絶対値のほうが大きければ，否定的な内容を表しているにも関わらず極性値の総和が正の値になり肯定的と認識してしまう．これを避けるため，肯定・否定を問う質問に対して単独で用いられた場合は総和が正の値になり，かつ WH 型の質問において他の語と共に使用される場合は “思う” 以外の語の極性値を反映させるため，経験的に 0.000001 と設定した．感情極性対応表における極性値は小数第 6 位までの実数で表現されているため，0.000001 は正の値で最小の実数となっている．

[3] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[4] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
[5] http://www.lr.pi.titech.ac.jp/~takamura/pndic_ja.html

3.2 発話文生成と重症度評価

この節では，発話文の生成とうつ病の重症度評価について述べる．本システムの発話文は，質問文，相槌，聞き直しの三種類からなる．前者二つの例は表 1 に記載した．重症度の評価は，ユーザからの応答を点数化することで行う．

まず，質問文と応答の点数化について説明する．QIDS は 16 の質問からなる．各質問には重症度の段階に対応した 4 つの選択肢があり，軽症のものから順に 0〜3 点が設定してある．被験者はその選択肢を選んでいくことで回答する．最終的に 16 の全質問の合計点を定められた方法によって調整した上で，0 点から 27 点の間で点数が確定する．この点数が 6 点以上の場合に「うつ病の可能性あり」と診断される．一定の点数幅毎に重症度が設定されており，点数幅と重症度の対応は表 2 で示したようになっている．

さて，各選択肢には定性的なものが含まれており，自由入力文から 4 択のどれに該当するかを適切に判定するのは困難である．そこで本システムでは，QIDS 各問の選択肢を質問文として改変・分割し，それぞれに当てはまるかを尋ねることで症状の程度を判断することとした．各質問から作った複数の質問文を 1 つの「質問セット」として考え，対話では同時に問いかけたほうが自然な一組の質問を統合し，15 の質問セットを作成した．各質問セットに含まれる質問文は，相当する重症度段階の低い順にユーザに対して提示する．感情極性評価型の質問の場合は応答の肯定・否定を分析した結果で，数値評価型の質問の場合は応答が QIDS が規定した値域に当てはまるかにより判定する．これが適合判定である限り次の段階でも繰り返し行い，不適合判定となった時点でその質問セットにおける評価を確定させ，次のセットへと移行する．このようにして，QIDS と同等の重症度評価を得る．質問セット毎の重症度判定フローの詳細を図 2 に示した．

相槌は，人手で作成されたテンプレートに基づいて，各質問セットにおいて重症度が確定した際に発話する．これを実施する目的は，ユーザに対して自然な対話感覚を与えるため，また診断に重要と思われる「共感」を表現するためである．相槌の生成時には，各質問セットに対して重症度ごとに用意したテンプレート文面を読み込み，ユーザの応答文書内の主語，動詞やその他の名詞等をテンプレート内の対応箇所に挿入する．

ユーザの応答を解析した際に，感情極性を伴う表現や数値が期待通りに得られなかった場合は，応答の聞き直しをする発話を行ってユーザに回答を換言するよう働きかける．聞き直しは最大で二度行い，それでも

表 2: QIDS における点数と重症度の関係

0 - 5 点	正常		16 - 20 点	重度
6 - 10 点	軽度		21 - 27 点	きわめて重度
11 - 15 点	中等度			

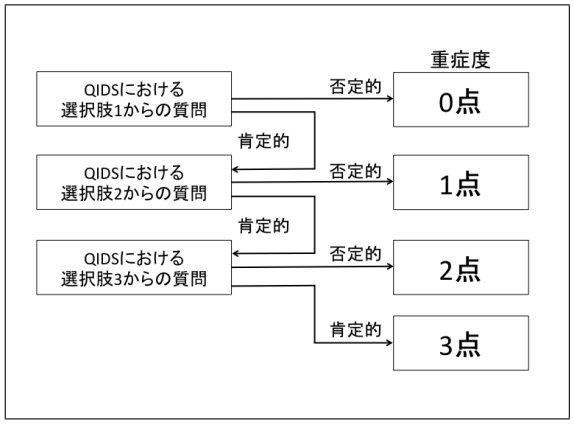

図 2: 質問セット毎の重症度判定フロー．ここでは全質問が肯定・否定を問う質問である場合を想定している．また，最初の質問に否定的な応答をした場合は，QIDS における重症度最低の選択肢を回答した場合に相当するため，重症度は 0 点となる．

有効な応答が得られない場合には，その質問の重症度判定は最低の値であると見なして，次の質問に進行する．一度目の聞き直しでは，ユーザに対して言い直してほしい旨の発言をする．二度目の聞き直しでは，肯定・否定がはっきりした回答，または数値を使った回答を促す旨の発言をする．

4 評価実験

4.1 実験データ

実験データとして，13 人の大学院生（男性 12 名，女性 1 名）から QIDS の結果および提案システムによる推定結果を収集した．

4.2 評価方法

推定した重症度の評価方法として，平均二乗誤差 (Mean squared error: MSE) とピアソンの相関係数 (correlation: r) を用いた．ここで，n は被験者数，x_i は被験者 i の QIDS の結果 (0 点から 27 点)，y_i は被験者 i の提案システムによる重症度推定結果 (0 点から 27 点)，$\bar{x}$ は全被験者の QIDS の結果の平均値，$\bar{y}$ は全被験者の提案システムによる重症度推定結果の平均値を表す．

表 3: 提案システムの推定精度

手法	MSE	r
提案システム	7.15	0.90

相関係数の計算方法は次の通りである．

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

平均二乗誤差の計算方法は次の通りである．

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \quad (2)$$

4.3 実験結果

提案システムの精度を表 3 に示す．提案システムによる推定結果は MSE=7.15，r=0.90 となり，高い精度で推定出来ることが確認出来た．

5 考察

提案システムでは，高い精度でユーザの精神状態を推定出来ることが確認出来た．これは，提案システムがユーザの自由記入文の意図を適切に判定出来ていることを示唆する．提案システムによる推定結果と QIDS による回答が異なった原因として多かったのは，システムが想定していない応答文が返ってきた場合である．例えば，体重の増減を聞いた時に「最近計っていない」または「変化していない」などの応答文が来た場合には上手く対応出来なかった．その他にも，打ち間違いを送信してしまったケースや，複雑な構造を持つ文章が送信されてきた場合に，ユーザの意図とは異なる解釈をするケースがあった．また，今回は大学院生を被験者として評価実験を行っているため，実際にうつ病に罹っている患者や，異なる年代の被験者が本システムを利用した場合，正しく結果を推定出来るかどうかを今後確認する必要がある．

6 おわりに

本研究では，ユーザとの双方向的な対話からユーザの重症度を判定するシステムを提案した．実験を行なった結果からは，高い精度で自己記入式の重症度判定シートの結果と近似することが確認出来た．この結果は，システムとの自然な対話から精神疾患の診断を適切に行える可能性を示唆するものであり，医療現場における診断補助や遠隔医療などといった分野での応用が期待される．

今後の課題としては，診断というドメイン外の対話にも対応させることが挙げられる．本システムでは，システムからの発話文は診断に関わる内容のみであり，ユーザからの応答文に対しても，診断に関する質問への回答のみが想定されていた．しかし，実際の臨床の現場では雑談や日常的な悩みの相談も行われる．このような対話に対応できるようにすることでより自然な形式での診断が可能になると考えている．

謝辞

本研究は奈良先端科学技術大学院大学情報科学研究科の 2013 年度 Creative and International Competitiveness Project の助成を受けて実施しました．心から感謝の意を表します．

参考文献

[1] François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research, Vol*, Vol. 30, pp. 457–501, 2007.

[2] Mizuki Morita, Soichi Ogishima, Kunihiro Nishimura, Eiji Aramaki, and Tateo Ito. Online population-based patient registry to collect and share health-related data of rare disease patients. In *AAAI Spring Symposium: Data Driven Wellness*, 2013.

[3] Arjun Mukherjee and Bing Liu. Improving gender classiffication of blog authors. In *In Proceeding of the 2010 Conference on Empirical Methods in Natural Language Processing. Assosiation for Computational Linguistics*, pp. 207–210, 2010.

[4] Philip Resnik, Anderson Garron, and Rebecca Resnik. Using topic modeling to improve prediction of neuroticism and depression in college students. In *In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1348–1353, 2013.

[5] Masoud Rouhizadeh, Emily Prud'hommeaux, Brian Roark, and Jan van Santen. Distributional semantic models for the evaluation of disordered language. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 709–714, 2013.

[6] 工藤真由美. アスペクト・テンス体系とテクスト —現代日本語の時間の表現—. ひつじ書房, 1995.

[7] 高村大也, 乾孝司, 奥村学. スピンモデルによる単語の感情極性抽出 (自然言語). 情報処理学会論文誌, 第 47 巻, pp. 627–637. 一般社団法人情報処理学会, 2006.