人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第19回)
SIG-AM-19-01

マルチタスク転移学習による
小規模教師データを用いた意図理解

Intention Understanding with Small Training Data Sets

by Utilizing Multi-Task Transfer Learning

城光英彰　内出隼人　小路悠介　大塚貴弘

Hideaki Joko, Hayato Ucihde, Yusuke Koji and Takahiro Ohtsuka

三菱電機株式会社　情報技術総合研究所
Information Technology R&D Center, Mitsubishi Electric Corporation

Abstract: In this research, we propose intention understanding method utilizing multi-task transfer learning. Our method improves intention understanding accuracy using data of different kind of domain as source domain. As source domain's training data, we use Japanese-English translation data (translation task) and Japanese Wikipedia data (sentence prediction task). As target domain's training data, we use transcribed utterance data of voice control of equipment. In this data, each utterance has one intention label. As an experimental result, we found that proposed method provides a performance improvement over previous transfer learning method in the case of small training data (the number of data for each intention label are 1, 3, 5, 10 and 30).

1. はじめに

機器の音声操作において，ユーザの発話文の意図を推定する意図理解は重要である．意図理解において Deep Neural Network (DNN) による方式の有効性が知られているものの [1]，適用分野（目標ドメイン）ごとに大規模教師データを作成する必要があり，コストがかかる問題が知られている．この問題に対し，大規模データが得られやすいドメイン（元ドメイン）のデータを活用し，目標ドメインで必要な教師データ数を削減する「転移学習」の有効性が報告されている [1]．筆者らも日英翻訳を元ドメインとした転移学習による意図理解方式を提案し，目標ドメインでのデータが小規模の場合の有効性を確認した [2]．この方式において，目標ドメインでの意図理解の正解率は，元ドメインのデータやタスクに強く依存しており，正解率を高めるためには，目標ドメインと「似ている」性質のドメインを元ドメインとすることが望ましいことが知られている [3]．しかし，このような性質のドメインの数は限られており，加えて，大規模教師データが手に入らない場合も多く問題である．

特定のドメインのデータやタスクに強く依存しない学習方式として，複数ドメインのデータやタスクを同時に一つの学習機で学習するマルチタスク学習がある [4]．これを元ドメインでの学習に利用することで，目標ドメインでの精度向上が狙える可能性がある．

目標ドメインでの精度向上のために，マルチタスク学習を元ドメインでの学習に利用した研究には，Subramanian et al. [5] のものがある．Subramanian et al. は，元ドメインにマルチタスク学習を用いた転移学習（マルチタスク転移学習）が，同義文判定タスクになどに有効であることを示した．しかし，Subramanian et al. が目標ドメインの学習に用いた教師データ数は同義文判定（二値分類）タスクに対し 1,000 から 25,000 データと，容易に収集できる量ではない．加えて，意図理解タスクでの検証もしていない．そこで，本研究では，マルチタスク転移学習による意図理解方式を提案し，教師データが小規模な場合の意図理解タスクでの有効性の検証をする．

提案方式に使用した学習機は Encoder-Decoder Model [6] である．元ドメインとして日英翻訳データ（日英翻訳タスク）および日本語 Wikipedia（文予測タスク）を，目標ドメインとして各発話文に一つの意図ラベルが付与されている機器操作データ（意図理解タスク）を用い，教師データ数と意図理解正解率の関係をもとめた．

2. 提案方式

提案方式は，日英翻訳タスクおよび文予測タスク（元ドメイン）により学習した DNN のパラメータを意図理解（目標ドメイン）に活用することによって少数データでも意図理解の精度を高める方式である．なお，文予測タスクとは，文書（本実験では日本語 Wikipedia）を文に分割したとき，文書のt番目の文s_t（対象文）から，次の文s_{t+1}と前の文s_{t-1}の予測をするタスク [7] である．

元ドメインおよび目標ドメインの学習に用いた DNN は，どちらも 1 層の Embedding Layer（300 次元）と 1 層の Hidden Layer（150 次元）からなる Bi-Directional Long Short Term Memory (Bi-LSTM) [8] を備えた Attention 構造 [9] を持つ Encoder-Decoder Model である（図 1）．転移学習の方式としては，元ドメインで学習したパラメータを，目標ドメインでの学習機の初期値として利用する INIT (Parameter Initialization) 方式 [1] を用いた. Embedding Layer は 300 次元，Hidden Layer は 150 次元である．

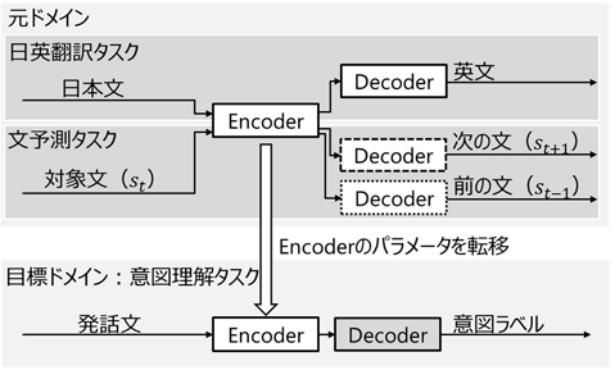

図 1: マルチタスク転移学習のイメージ図. 元ドメインとして日英翻訳データ（日英翻訳タスク）および日本語 Wikipedia（文予測タスク）を，目標ドメインとして各発話文に一つの意図ラベルが付与されている機器操作データ（意図理解タスク）を用いた．

表 1: 使用した教師データ

教師データ	使用ドメイン	教師データ数	入力文の語彙数
Tanaka Corpus	元ドメイン	25,176 対	8,875
日本語 Wikipedia	元ドメイン	21,791 対（次の文） 21,703 対（前の文）	20,249
機器操作データ	目標ドメイン	5,600 対	20,247

3. 実験

3.1. 使用データ

元ドメインの教師データには，日英翻訳データとして Tanaka Corpus[1]を，日本語 Wikipedia として，Wikipedia 日本語版のダンプ[2]を用いた．目標ドメインとして各発話文に一つの意図ラベルが付与されている機器操作データを用いた（表 1）．意図ラベルの種類は 56 種類であり，各意図ラベルには 100 個の発話文が対応している ．前処理として，Tanaka Corpus の，意味的にあいまいな対，文字長が 100 以上の対の修正や削除をした. また，計算時間削減の目的で，Tanaka Corpus と日本語 Wikipedia ともに，入力文 の単語の 8 割が，機器操作データに出現する単語で構成されているもののみを，教師データとして使用した．

表 2: 目標ドメインで使用した教師データ数: 合計の教師データ数は「意図の数」と「各意図ラベルに対する教師データ数」の乗算値である

意図の数	各意図ラベルに対する教師データ数	合計の教師データ数
56	1	56
	3	168
	5	280
	10	560
	30	1,680
	45	2,520
	90	5,040

3.2. 実験内容

実験条件を表 3 に示す．パラメータの転移手法は Embedding Layer のパラメータを転移するもの（EMB），Embedding Layer と Hidden Layer を転移するもの（ALL）の二種類である．元ドメインのタスクとしては，翻訳，文予測，翻訳と文予測を同時に行うマルチタスクの三種類がある．そのため，実験条件は合計で六種類となる．

実験内容を次に示す．

実験１：まず，教師データ数を 10 としたときの各実験条件の意図理解正解率を算出し，提案方式が小規模教師データにおいて有効であることを示す．

実験２：次に，従来方式と提案方式の各々の中でもっとも正解率が高い実験条件について，意図ラベルごとの教師データ数を 1, 3, 5, 10, 30, 45, 90 と変化させたときの意図理解正解率を算出し，教師データが小規模なほど，提案方式が有効なことを示す．

実験１，実験２ともに，正解率は 10 分割交差検定

[1] http://www.edrdg.org/wiki/index.php/Tanaka_Corpus
[2] https://dumps.wikimedia.org/jawiki/, 2017 年 11 月取得

により算出する．また，算出結果の信頼性を高めるため，各交差検定について，評価用のデータ 10 個以外の全てのデータ 90 個を少なくとも一回は教師データとして使用するよう複数回の学習をする．具体的には，意図ラベルごとの教師データ数が 1 のときは 90 回，3 のときは 30 回，5 のときは 18 回，10 のときは 9 回，30 のときは 3 回，45 のときは 2 回，90 のときは 1 回の学習をする．つまり，例えば教師データ数が 10 の場合は，合計 90 回（9×10）の学習および意図理解正解率の算出をすることになる．

表 3: 元ドメインにおける学習の実験条件．EMB は Encoder の Embedding Layer のパラメータの転移を，ALL は Encoder の全てのパラメータの転移を表す．

実験条件の名称	元ドメインの教師データ		転移するパラメータ	
	日英翻訳データ	日本語 Wikipedia	Embedding Layer	Hidden Layer
従来方式: 翻訳（EMB）	使用	-	転移	-
従来方式: 翻訳（ALL）				転移
提案方式:マルチタスク（EMB）	使用	使用	転移	-
提案方式:マルチタスク（ALL）				転移
文予測（EMB）	-	使用	転移	-
文予測（ALL）				転移
転移学習なし	-	-	-	-

3.3. 実験結果

実験 1：まず，教師データ数を 10 としたときの各実験条件の意図理解正解率をもとめた．結果を図 2 に示す．全体として，Embedding Layer のみの転移（EMB）の方が，Encoder の全てのパラメータを転移した場合（ALL）よりも正解率が高いことがわかる．また，全実験条件の中でもっとも正解率が高い実験条件は「マルチタスク（EMB）」であることがわかる．この，「マルチタスク（EMB）」と，各他実験条件の意図理解正解率について，等分散の仮定の下で右片側二標本 t 検定を適用したところ，「マルチタスク（EMB）」の正解率が，各他実験条件と比べ有意に高いことが確認できた（有意水準$\alpha = 0.05$）．

実験 2：次に，従来方式と提案方式の各々の中でもっとも正解率が高い実験条件である，「翻訳（EMB）」と「マルチタスク（EMB）」について，意図ラベルごとの教師データ数を 1, 3, 5, 10, 30, 45, 90 と変化させたときの意図理解正解率をもとめた．結果を図 3 に示す．各意図ラベルに対する教師データ数が 1, 3, 5, 10, 30 と少ないとき，提案方式の意図理解正解率は従来方式と比較し高くなることがわかる．この結果について実験 1 と同様の条件で検定を行ったところ，教師データ数が 1, 3, 5, 10, 30 と少ないとき，「マルチタスク（EMB）」の正解率が，「翻訳（EMB）」と比べ有意に高いことが確認できた．

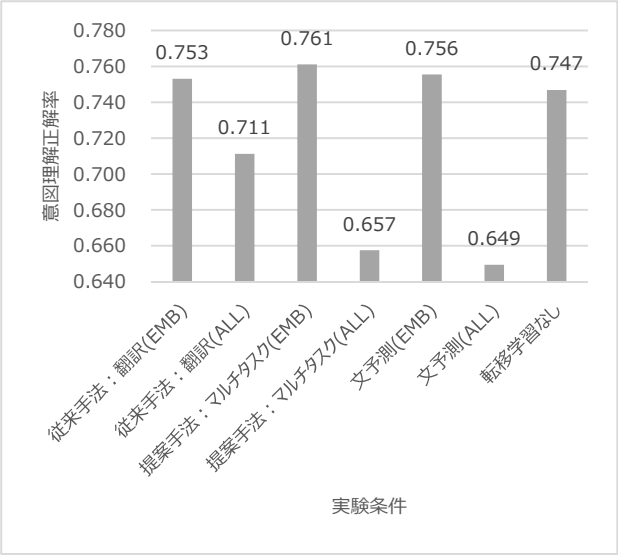

図 2: 教師データ数を 10 としたときの，各実験条件の意図理解正解率．「提案方式：マルチタスク（EMB)」の正解率が，他手法と比べ有意に高い．

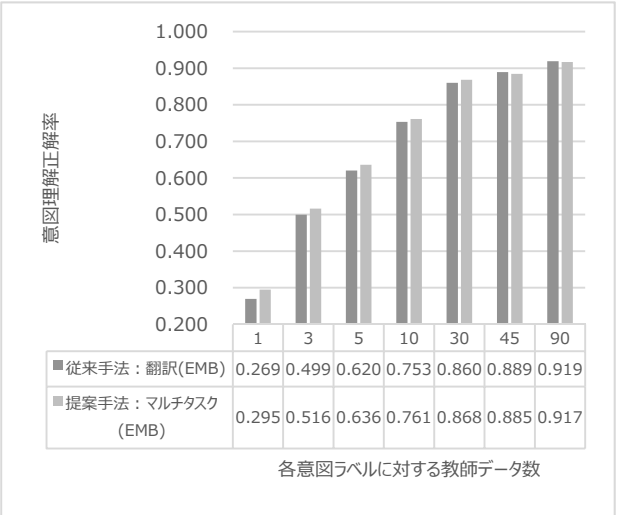

図 3: 従来方式と提案方式における目標ドメインの教師データ数を変化させたときの意図理解正解率．各意図ラベルに対する教師データ数が 1〜30 と少ないとき，提案方式の意図理解正解率は従来方式と比較し有意に高い．

また，提案方式は教師データ数が少ないほど有効なこともわかった．特に，各意図に対する教師データ数が 1 の場合では, 正答率は従来手法「翻訳（EMB）」の 0.269 から提案手法「マルチタスク（EMB）」の 0.295 へと，2.6 ポイント向上した．この向上の理由は，元ドメインにマルチタスク学習を適用したことにより，元ドメインにおいて，特定のドメインのデータやタスクに強く依存しない学習ができたからと考えられる．

4. まとめ

本研究では，小規模教師データを用いたマルチタスク転移学習方式を提案し，意図理解正解率の評価を行った．元ドメインとして日英翻訳データ（日英翻訳タスク）および日本語 Wikipedia（文予測タスク）を，目標ドメインとして各発話文に一つの意図ラベルが付与されている機器操作データ（意図理解タスク）を用い，教師データ数と意図理解正解率の関係をもとめた．その結果，各意図ラベルに対する教師データ数が 1, 3, 5, 10, 30 と少ないとき，提案方式の「マルチタスク（EMB）」が，従来方式である「翻訳（EMB）」の意図理解正解率を有意に上回り，目標ドメインでの教師データ数が少ない場合において提案方式が有効であることがわかった．今後は，元ドメインのデータおよびタスクを追加することで，意図理解正解率の向上を目指す．また，目標ドメインのタスクを追加し，元ドメインで学習したパラメータの汎用性の評価も行う予定である．

参考文献

[1] Mou, L., Meng, Z., Yan, R., Li, G., Xu, Y., Zhang, L., and Jin, Z.: How transferable are neural networks in nlp applications?, In EMNLP, (2016)

[2] 城光英彰，内出隼人，小路悠介，大塚貴弘: 転移学習による小規模教師データを用いた意図理解，電子情報通信学会 全国大会, (2018)

[3] 神嶌敏弘: 転移学習，人工知能学会誌, Vol. 25, No. 4, pp. 572-580, (2010)

[4] Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O., and Kaiser, L.: Multi-task sequence to sequence learning, In International Conference on Learning Representations, (2015)

[5] Subramanian, S., Trischler, A., Bengio, Y., and Pal, C. J.: Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning, In ICLR, (2018)

[6] Sutskever, I., Vinyals, O., and Le, Q. V.: Sequence to sequence learning with neural networks, In Advances in neural information processing systems, pp. 3104-3112, (2014)

[7] Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S.: Skip-thought vectors, In Advances in neural information processing systems, pp. 3294-3302, (2015)

[8] Ma, X. and Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNsCRF, In ACL, (2016)

[9] Bahdanau, D., Cho, K. and Bengio, Y.: Neural machine translation by jointly learning to align and translate, In International Conference on Learning Representations, (2015)

日本語学習者会話データベースにおける 隣接ペアの談話促進に関する一考察
～対話破綻検出や対話システムへの適用を目指して～

A Study on Promotion of Discourse of Language Action and Adjacent Pair in Japanese Learner's Conversation Database

○ 太田 博三[1]

Hiromitsu Ota 1[1]

[1]放送大学 教養学部
[1]The Open University of Japan

Abstract: In recent years, interactive systems and dialogue generation in natural language processing have attracted attention. Due to the spread of the chat bot to the call center, an accurate human interactive response is required. On the other hand, qualitative interactions in sociology's ethnomethodology and discourse analysis / conversation analysis are beneficial. Therefore, once again, using the Japanese language learner conversation data corpus provided by the National Institute for Japanese Language, we examine the effect and aim at applying to the tendency of dialogue breakdown and dialogue generation.

1. はじめに

1.1 研究の背景と目的

スマートスピーカーが家庭に普及し，自動運転が実用化されようとしている中，従来から発展し商用化されているロボットの Pepper や各種チャットボット（Chatbot）は，人間と比べて，小さくない乖離があると指摘され，4 年以上が経過している．チャットボットのコールセンター等への導入も 2 年以上経過している．ここでは，制御文による対話応答が第 1 義的に実装され，第 2 義的にディープラーニング（Deep Learning）による運用が，もくされた．しかし，いずれも不完全燃焼に終始している．どちらか一方，もしくは折衷でも，人間に代替する品質に昇華できていないと考え，調べ始めたのが本考察のきっかけである．

次に，エスノメソドロジーや会話分析の勉強会に参加した際，対話システムのような粗い応答では不十分だという趣旨のご発言と勝手ながら解釈した機会があり，追及してみようと考えた．

そこで，現行の対話システムに定性的な視点で考察し，定量的な分析に持ち込むことで，スケール化させ，実用化に結びつける第一歩にしたいというのが目的である．全体的にディープラーニングに適用したらよいか，部分的かも検討したい．

1.2 研究の新規性

本研究の遠い新規性となるが，もっぱら，数量データによるディープラーニングに定性的な要素を取り入れたいという点であるが，本稿では，誰もが入手可能なデータである日本語学習者の会話データを用いることで，統計的な有意性やサンプル数より，日常生活の感覚でわかることを重視したものである．次に，[質問]-[応答]や[申し出]-[受諾/拒否]などの隣接ペアの類型が上記のデータにどのくらいあてはまるかなど，計量化してみた．ここで，実証的な知見が得られれば，話者交代の予測や対話の破綻をしても修復する発話を学習させるなど，次につなげられる．具体的には，隣接ペアの次には，主に以下の 5 つが考えられる．

1) Yes/No の応答詞 ：あー，うん，えー，そう
2) あいづち ：んー，はい，はー，えー
3) 言いよどみ：んー，あー，えー
4) 呼びかけ ：ね，ねー
5) フィラー ：あのー，そのー，えーと，えっと

今後，これらを分析し追加することで，更なる対話システムの質的向上につながる可能性があると考えられる．

1.3 研究の主な手法

基礎集計を中心に行いながら考察する．国立国語研究所の提供している「日本語学習者会話データベース」を用いて集計を行う．隣接ペアは本稿で定義する種類のものに限定し計量化する．次に，それらのペアが全体の会話の促進になっているかなどを考察する．また，その隣接ペアの前後，もしくは直後の発話が修復に向けてのものか，完全に破綻しているが強引に会話を続けたものであるのかも含めて，定性的な判断を行う．

1.4 用いたデータセットについて

国立国語研究所が公開しているコーパスの中の1つである「日本語学習者会話データベース」（図1）を用いる．またKYコーパスも同様の趣旨で作られたものであり，適宜，用いた．1990年の入管法の改正により，日本の社会状況に応じて，外国人受入れの適切な方策が必要となり，日本語学習を必要とする住民（言語生活者）の需要に見合った言語教育の展開が期待されていた．ACTFL-OPI（全米外国語教育協会認定の面接式口頭能力テスト）を活用し，日本語を用いた自然な会話に限りなく近い対話で構成されている．

図 1. 属性別の日本語教育会話データベースの検索画面

ACTFL（全米外国語協会）によるOPI(Oral Proficiency Interview Test)に基づいており，日本語OPIは1993年に発足し，15年近く経過している．ここでの判断尺度は，次の4つに区分されている．

1) 超級（Superior）
2) 上級（Advanced）
3) 中級（Intermediate）
4) 初級（Novice）

これは「日本語学習者会話データベースの利用手引き（平成 22 年 5 月 国立国語研究所）」によれば，言語運用能力は10種類の階級に区分されている(表1)．対話のスクリプトは，インフォーマント（日本語学習者/データ提供者）とテスター（面接者）とからなり，30分ほどの対話形式で構成されている．

また上記の10段階のOPIレベルや性別，年齢，出身国などを選択することができる．検索条件を設定してダウンロードすると，文字化（一部，音声化）されたスクリプトが入手でき，有用である．

表 1-. OPI 能力区分表

区分	OPIレベル	階級	OPI評価
1	超級（Superior）	1	超級
2	上級（Advanced）	2	上級−上
	〃	3	上級−中
	〃	4	上級−下
3	中級（Intermediate）	5	中級−上
	〃	6	中級−中
	〃	7	中級−下
4	初級（Novice）	8	初級−上
	〃	9	初級−中
	〃	10	初級−下

2. 先行研究

本考察では，下記の3つ区分した．1つ目は，エスノメソドロジーや会話分析などの社会学である．言語学も多分に含まれている．2つ目は，対話システムを支える自然言語処理，3つ目は，深層学習，すなわちディープラーニングである．

2.1 エスノメソドロジー・会話分析

坊農・高梨他(2009)では，隣接ペアとは，[質問]−[応答]の対をなす発話の連鎖を指すものとして，対話システムにおける対話モデルに発話連鎖構造の土台としているとある．さらに，隣接ペアの概念には，[質問]に対し，[応答]がなされなかった場合に，どのような修復連鎖や挿入連鎖構造が生起しながら会話が進行するかを述べている．魏(2015)は「あのー」や「まー」などをフィラーと定義し，発話者が何らかの心的操作を行っている最中に発するもので，場をつなぐ機能を持つ言葉と定義している．多くは「感動詞」や「間投詞」に区分される．このフィラーを使いこなすのも，あいづちなどと同じく，会話をつなぐ言葉として，留意したいと考えている．

2.2 自然言語処理

対話システムに実装される可能性は示している．また，徳永・乾・松本(2005)及び徳永(2014)は，チャット対話の収集からコーパス作成，そしてチャット対話の構造モデルを提案している．このチャット対話の質問や返答などの談話機能を担う構成単位が交換行為である．交換単位は「働きかけ」，「応答」，「補足」の3種類に区分され，さらに2, 3の枝葉に分かれている．また，素性に関する考察は有益であり（表2），本研究ではこれらを精緻化することが具体的な目標でもある．素性の組合せと継続関係の同定や再現率は2人の場合でも3人の場合でも，86%と高く，素性も厳選されている．発言間の結束度は次の式で求めている．< n(名詞)，rel(助詞)，v(動詞)>の共起確率 $P(<n, rel, v>)$を求める．この確率 $P(<n, rel, v>)$ は，Probabilistic Latent Semantic Indexing(PLSI)で推定する．単語の共起を潜在的な意味から同時発生とみなす手法である．PLSIにおける共起確率 $P(<n, rel, v>)$は次の式で与えられる．

$$P(<n, rel, v>) = \Sigma P(<rel, v|z)\ P<rel, v|z)P(n|z)P(z).$$

ここで，zは共起の潜在的な意味クラス(隠れクラス)を指しており，パラメータの $P(<rel, v|z>)$，$p(n|z)$，$p(z)$)はEMアルゴリズムで推定している．

表2 素性一覧徳永・乾・松本(2005)

素性	素性の説明
発言の末尾の表層表現	各発話の末尾が句点，読点，クエスチョンマークであるか否かの2値
CRRuとPREu間の発言時間の差	CRRuとPREu間の発言時間の差が2分以上であるか否かの2値
発言間の結束度	共起確率に基づくCRRuとPREu間とCRRuとNBNu_s間の結束度の強い方を1とする2値
交換行為の対話クラス	対話行為辞典に各交換行為(20種類 ≒ 隣接対)のクラスに分類したもの
交換行為の末尾の表層表現	同一人物の複数読み取れる発話の一番最後の末尾がクエスチョンマークであるか否かの2値
交換行為の発言時間の差	CRRuとPREmの先頭の発話における発話時間の差が5分以上であるか否かの2値

2.3 対話自動生成のディープラーニング

対話応答の自動生成に関しては，ICML Workshop(2015)でVinyals et al(2015)のGoogleのチームがNIPS2014で発表されたSequence to Sequence modelを基としている．多層のLong-Short term memory(LSTM)を用いて文章をベクトル化（エンコード）し，別の多層LSTMを用いてベクトルをデコード（復元）するものである．これは「日本語ー英語」間の機械翻訳でよく用いられているアーキテクトであり，従来と比べて，自然な会話を生成するようになった． Ghazvininejad et al(2018)は，上記のモデルを拡張発展させたものである．会話型だけでなく非会話型データも組み合わせることによりSeq2seqにおけるNeural Conversation Modelを発展させたものである．

2.4 対話システムと会話ユーザーインターフェース

狩野(2017)では，現在に至るまでの対話システムと将来の展望を簡潔にまとめられている．1960年代に開発されたELIZAや人工無能から，「○」「△」「×」などのアノテーションによるチューリングテストを経て，現代の雑談対話システムの1つである2016年に発表された論文に基づくMicrosoft社の「りんな」まで網羅している．「りんな」では，発話ペアデータと教師付き機械学習は統計的な対話システムの多くに共通していることが少なくない．また，対話データを正解データとしてくんれんする強化学習では，状態遷移の訓練になるため，会話の流れを学習することになり，未知の対話に対応することが期待されている．

2.4 国内外での取り組み

対話破綻検出チャレンジ(2015-2017)やDTFC7, NTCIR-14など，年次でハッカソンのような国際的な大会として開催され，集合知となっている．

3. 基礎集計と分析による考察

日本語学習者会話データベース全体的にデータを見渡してみると，全データは390個ある．インフォーマント(日本語学習者)の属性は，20代が圧倒的に多く，女性が男性の2倍近くおり，大半を占めている．日本語学校生や大学・大学院生が半分を占めている（図1）．

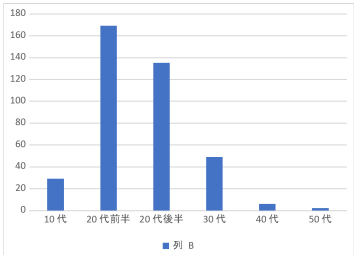

図3.1 インフォーマントの年代別分布

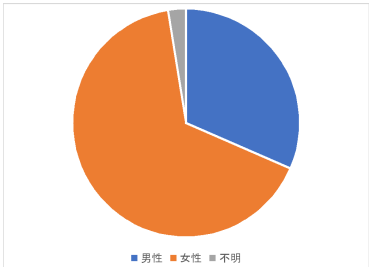

図3.2 インフォーマントの性別

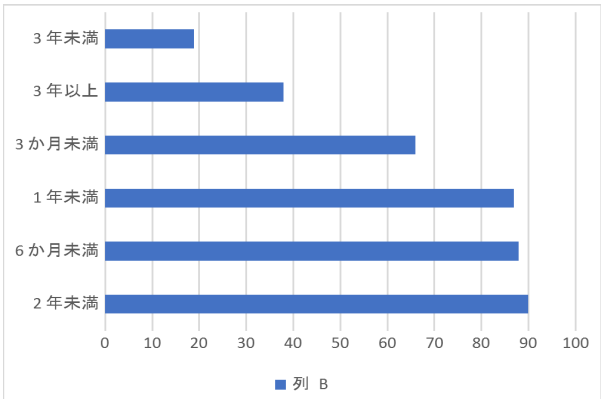

図 3.3 インフォーマントの日本滞在時間

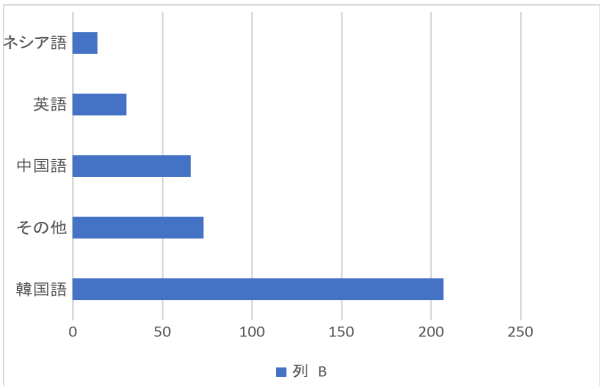

図 3.4 インフォーマントの日本語学習期間

3.1 インフォーマント属性間の比較について

本考察では，自然な対話文の自動生成を目指していることから，次の 2 つを比較考察する．OPI の判断尺度は，国立国語研究所(2010)「日本語学習者会話データベースの利用手引き」に準じ，超級と中級とで比較考察した．超級は人間と仮定し，中級はチャットボットなどの機械としてみた．主な選択した要因は次の 2 つである．

1) 流暢さ
2) 語用論的能力

超級では，「流暢さ」とは，会話全体がなめらかであること，これに対して，中級では，「流暢さ」とは，つかえることが多く一人で話つづけるのは難しいと定義されている．一方，超級の「語用論的能力」とは，ターンテイキングや間の取り方，相づちなどが巧みにできると定義されているのに対して，中級は，相づちや言い換えなどに成功するのはまれとされている．

3.2 超級と中級のデータについて

中級は年齢が 20 代半ばとまとまって分布しているのに対し，超級は 23 歳から 49 歳とばらつきが大きい．これはデータ数が 9 つと少ないためであるが，出身国と母国語は超級のハンガリーのインフォーマントを除いて，韓国である．滞在期間が大きく異なり，
超級は 5-10 年が大半であるのに対し，中級は 3 か月から 6 か月の間に分布している．

表 3.1 超級のインフォーマント属性

	データ番号	OPIレベル	年齢	性別	出身国	母語	職業等	日本滞在期間	日本語学習期間(参考)	日本語能力試験(参考)
1	10	超級	28	女	韓国	韓国語	会社員	5年	7年	－
2	15	超級	26	男	韓国	韓国語	専門学校生	3年	18年	－
3	76	超級	34	女	韓国	韓国語	主婦	5年	4年	－
4	202	超級	35	女	韓国	韓国語	講師	10年	6年	1級
5	230	超級	26	男	中国	中国語	大学院生	5年1ヶ月	7年	－
6	258	超級	43	女	ブルガリア	ブルガリア語	教師	18年	22年？	－
7	323	超級	23	男	韓国	韓国語	大学生	5年	8年	－
8	338	超級	49	男	韓国	韓国語	会社員	22年	2年	1級
9	349	超級	40	女	韓国	韓国語	大学教員	15年	2年～3年	－

表 3.2 中級のインフォーマント属性

	データ番号	OPIレベル	年齢	性別	出身国	母語	職業等	日本滞在期間	日本語学習期間(参考)	日本語能力試験(参考)
1	2	中級-下	22	女	韓国	韓国語	大学生	3ヶ月	11ヶ月	－
2	12	中級-下	25	男	韓国	韓国語	日本語学校生	5ヶ月	5ヶ月	－
3	26	中級-下	27	女	韓国	韓国語	日本語学校生	5ヶ月	18ヶ月	－
4	8	中級-中	24	女	韓国	韓国語	日本語学校生	3ヶ月	7ヶ月	－
5	9	中級-中	25	女	韓国	韓国語	日本語学校生	2ヶ月	8ヶ月	－
6	22	中級-中	28	女	韓国	韓国語	日本語学校生	5ヶ月	1年	－
7	6	中級-上	24	女	韓国	韓国語	日本語学校生	6ヶ月	1年6ヶ月	－
8	7	中級-上	28	女	韓国	韓国語	専門学校生	2年	23ヶ月	－
9	11	中級-上	26	女	韓国	韓国語	日本語学校生	6ヶ月	9ヶ月	－

表 3.3 比較に用いたデータセット数

OPIレベル	母数	使用したデータ数	合計
超級	9	9	9
中級-下	36	3	
中級-中	84	3	9
中級-上	68	3	

3.3 隣接ペアとその計量化の検討

隣接ペアの重要な特性に，第 1 部分(First-Pair-Part: FPP)が産出されると，それに対応する特定の型の第 2 部分(Second-Part-Pair: SPP)の産出が条件的に適切になると前川・小磯他(2015)は言及している．本節では，試みの一環として，形態素解析した後に，同じ語句がでてきたら，その合計の半分として数量化した後に，目視で確認をすることにした(表 3.3.1，表 3.3.2，表 3.3.3)．隣接ペアが見出しやすい次の 4 つの品詞に焦点を当てて考察することにした．

1) 名詞
2) 感動詞
3) 間投詞
4) 応答詞

表 3. 3. 1

隣接ペアである例		
‘chiba-1232:514.5590-516.4996		
A1:	選べんだ	←第1部分(FPP)
B2:	選べる	←第2部分(SPP)
A3:	へえ	

表 3.3.2

隣接ペアでない例		
‘chiba-0332:437.2296-441.4541		
C1:	私も動物飼いたいな:	←働きかけ(I)
C2:	植物でもいいや	←働きかけ(I)
A3:	うん	←応答(R)

中級データセット
1) 頻出名詞(上位 10 件)
ん(2254) ー(1717) 笑(649) 私(294) こと(268) 音(249) 人(247) 息(246) 日本(183) お(169) 今(168) , (168) 韓国(167)

1) 頻出詞(上位 10 件)
2)感動詞
はい(2594) あー(944) あ(530) えー(309) そうですね(61) ありがとう(55) はーい(52) え(48) ま(41) ふーん(33) うん(30)

超級データセット
1)名詞
ん(2869) ー(1778) 笑(546) の(376) こと(342) 人(310) それ(292) 日本(261) 今(199) 韓国(184)

2)感動詞
はい(2114) えー(1038) あー(389) あ(273) そうですね(183) ま(159) なるほど(97) え(89) ふん(65) ふーん(54)

3.4 結果の考察
中級の名詞では，ん(2254) ー(1717) 笑(649) 私(294)などが多く，主観性が見受けられた．その一方で，超級の名詞では，人(310) それ(292) 日本(261) 今(199) 韓国(184)などのようぬ，代名詞や国柄を表す語句が見受けられ，客観性が見受けられた．
また，中級の感動詞では，はい(2594) あー(944) あ(530) えー(309) はーい(52)など，あいまいさが見受けられた．一方で，超級の感動詞では，はい(2114) えー(1038) そうですね(183) ま(159) なるほど(97) など確かな返答が見受けられた．

4. 今後の展望

本研究はチャットボットなどの対話システムを対象としたものであるが，今後は次のような視点で，ロボテックスを対象とした研究につなげたい．秋谷・丹波・久野・山崎他(2007)では，介護ロボットの実現に向けて，介護者と高齢者との相互行為を深く分析したものである．今後はロボットに搭載されることが予見される．相互行為の視点が，より人間的になると考えられ，期待されている．

謝辞

本研究の一部は，学部時代にマレーシア語や生成変形文法を教えて頂いた正保勇先生(東京外国語大学名誉教授)の雑談の中での教えが大きく影響している。また計量社会科学に関しては，博士課程時代にご指導頂いた聖学院大学大学院の松原望先生(東京大学名誉教授)を想起しながら試行錯誤できた．ここに謝意を表したい．

参考文献

[1] 国立国語研究所(2009) 「日本語教育ネットワーク」https://nknet.ninjal.ac.jp/nknet/ndata/opi/
[2] 国立国語研究所(2010)「日本語学習者会話データベースの利用手引き」
[3] 鎌田・山内「タグ付き KY コーパス」 http://jhlee.sakura.ne.jp/kyc/corpus/
[4] 坊農・高梨他(2009)「知の科学 多人数インタラクションの分析手法」3 章, 人工知能学会編集, オーム社
[5] 徳永・乾・松本(2005)「チャット対話における発言間の継続関係と応答関係の同定」 自然言語処理 言語処理学会
[6] 徳永(2004)「チャット対話における発言の継続関係と応答関係の同定」修士論文 奈良先端科学技術大学院大学 情報科学研究科 情報処理学専攻
[7] 牧野成一他(2001)「ACTFL-OPI 入門」アルク
[8] Vinyals,et al(2015) Quoc Le. A Neural Conversational Model, arXiv
[9] Ghazvininejad(2018) A Knowledge-Grounded Neural Conversation Model. Microsoft
[10] 喜連川他(2017)「暗黙の発話状況を考慮したニューラル対話モデル」. 言語処理学会 第 23 回年次大会 発表論文集
[11] 串田, 平本, 林(2017)「会話分析入門」 勁草書房
[12] 対話破綻検出チャレンジ 2015 https://sites.google.com/site/dialoguebreakdowndetection/
[13] 船越・東中他(2016)「対話破綻検出チャレンジにおける対話破綻データと破綻検出結果の分析」 言語処理学会 第 22 回年次大会
[14] 東中・船越(2016) Project Next NLP 対話タスクにおける雑談対話データの収集と対話破綻アノテーション 人工知能学会 SLUD
[15] DSTC7(2017) Dialog System Technology Challenges http://workshop.colips.org/dstc7/index.html
[16] NTCIR-14(2018-19) - 国立情報学研究所

http://research.nii.ac.jp/ntcir/ntcir-14/index-ja.html
[１７] EMCA 研究会 エスノメソドロジー・会話分析とはなにか - http://emca.jp/learn
[１８] 狩野(2017) 「コンピューターに話が通じるか対話システムの現在」 情報管理 Vol.59 no.10
[１９] 石崎・伝(2001)「談話と対話」東京大学出版会
[２０] 藤田 自然言語表現の言い換え http://paraphrasing.org/~fujita/paraphrasing-ja.html
[２１] 魏(2015)談話におけるフィラー「ま（一）」の待遇差に関する予備的考察, 山口大学東アジア研究学術雑誌論文
[２２] Smith et.al(2000) Conversation Trees and Threaded Chats, CSCW
[２３] 秋谷・丹波・久野・山崎他(2007)「介護ロボット開発に向けた高齢者養護施設における相互行為の社会学的分析」電子情報通信学会論文誌 D Vol.J90-D No.3 pp. 798-807

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第19回)
SIG-AM-19-03

グループを対象とした合議不要な観光スポット推薦手法の提案

Proposal of Sightseeing Spot Recommendation for Group without Discussion

秦 馳[1]　高間 康史[1]

Chi Qin[1] and Yasufumi Takama[1]

[1]首都大学東京大学院システムデザイン研究科
[1] Graduate School of System Design, Tokyo Metropolitan University

Abstract: This paper proposes a method of recommending sightseeing spots without discussion for a tourist group. As members in a group usually have different interests in sightseeing spots, it tends to take a lot of time to decide a sightseeing plan which satisfies all members' preference. Furthermore, it is difficult for those who are not good at expressing their opinions to take part in the discussion. With the proposed method, each member in a group inputs his/her interests and conditions about sightseeing spots, and then evaluates the recommended spot list one by one. This paper also proposes to determine the order of evaluating the list on the basis of the questionnaire using the MBTI taxonomy, which have been proposed in the field of psychology. Effectiveness of the proposed method is shown by user experiment.

1. はじめに

本稿では，グループで旅行する際に訪問する観光スポットを効率よく決定可能とするために，合議不要な観光スポット推薦手法を提案する．

近年，週末や休暇などを利用して旅行をする人は増加している．観光先を選択する際には，観光したい都道府県の観光サイトや旅行会社の予約サイトなどを使って，観光スポットの紹介文，写真や口コミを探して参考にする人も多い．しかし，ウェブサイトのデータ量は膨大になっており，その中から興味がある観光スポットを発見することは困難となってきている．このような背景から，観光スポット推薦に関する研究が行なわれている[1].

これらの観光推薦システムは個人ユーザを対象としたものが多いが，友人や家族などグループで観光する場合も多く存在する. グループで観光する場合，嗜好や興味などが互いに異なることが多く，全員が満足する計画を合議で決定することは一般に難しく，時間もかかる．また，グループの人数が多い場合は特に，全員が集まって議論することが困難な場合も多い．さらに，討論が苦手な人の意見が反映されにくいという問題点も存在している．

これらの問題を解決するために，本稿では，グループでの観光計画を想定し，合議不要な観光スポット推薦を可能にする手法を提案する. 提案手法では，最初にグループのメンバー各自に観光スポットに関する興味や条件を入力してもらう．その後，グループ全員が入力した情報に基づき決定した推薦観光スポットのリストについて，一人ずつ順番に確認・評価をしてもらう．各メンバーは，自分が気になっている観光スポットをリストから選択する他，行きたくないスポットをリストから削除する．評価した結果に基づき，次のメンバーに提示する推薦スポットリストが更新される．最後のメンバーによる確認・評価が終わった後，グループに対する推薦スポットリストが得られる．

提案手法において，推薦スポットリストを確認する順番が重要である．本稿では，MBTI (Myers-Briggs Type Indicator) 分類法[2]を利用したアンケート結果に基づき，メンバーの確認順序を決定する手法も提案する．ユーザ実験を行い，提案手法の有効性を示す.

2. 関連研究

2.1　観光スポット推薦手法

鈴木らは個人の嗜好を分析し，地域の特性も考慮して訪問推奨エリアを推定・可視化するシステムを提案している[3]．嶋田らはユーザに好きな観光スポット等を入力してもらい，それに類似する観光スポ

ットを推薦するシステムを提案している[4]．奥薗らは複雑な操作なしに複数人の嗜好を反映可能な観光地推薦システムを提案している[5]．このシステムでは，グループの嗜好を分析するために，旅行に関するイメージを各ユーザに提示し，AHP (Analytics Hierarchy Process) に基づく見解距離均等法を用いてユーザの嗜好を統合している．ユーザはイメージの選択を行うだけなので気軽に利用可能と考えられるが，観光スポットに対する具体的な条件や興味を扱うことは困難と考える．

2.2 MBTI 分類法

MBTI (Myers-Briggs Type Indicator) は，Jung の心理学的類型論 (Psychological Types) [6]をもとに，Myers によって研究開発された自己理解メソッドである．他者との違いを知ってお互いに尊重しあうことを目的に作成されている．MBTI はカウンセリングやコンサルティング，人事教育などで活用されている[7,8,9].

MBTI は，Jung の類型論の指標に判断の態度 (J: Judging) と知覚的態度 (P: Perceiving) という独自の指標を加えて, 以下の 4 指標に基づき 16 タイプに性格を分類する．

・内向 (I: introversion) －外向 (E: extraversion)
・感覚 (S: sensing) －直感 (N: intuition)
・思考 (T: thinking) －感情 (F: feeling)
・判断 (J: judging) －知覚 (P: perceiving)

MBTI では 16 種類の性格を役割，および戦略の 2 層で構造化している．役割の層は，目標，興味，優先行動を表し，分析家，外交官，番人，探検家に分類される．

分析家（直感的論理型：NT）は，合理的で公平な思考の持ち主で，知的な討論に長けている．実用的な視点を持ち，皆を満足させるものよりも機能的なものに興味を持っている．

外交官（直感的道理型：NF）は，共感と協力を重視し，集団において調整役になることが多い．

番人（現実的計画型：SJ ）は，協調性があり，現実的な思考の持ち主で，秩序，安全，安定を重視している．特に階級や規則が明確な環境の中で力を発揮する．

探検家（現実的調査型：SP ）は，最も自発的であり，実用的で現実的な思考に基づき，迅速な判断・行動が要求される状況で力を発揮する．

3. 提案手法

提案手法では，以下の手順に従いグループに対する推薦スポットリストを生成する．

(1) メンバーごとに, 観光スポットに対する興味・条件を入力する．
(2) 推薦スポットリストを生成する．
(3) 推薦スポットリストを確認するメンバーの順番を決定する．
(4) 順番に従い一人ずつ推薦スポットリストを確認する．

(1) は，グループメンバーそれぞれに興味のある観光スポットの特徴，知名度，コストと滞在時間を回答してもらう．観光スポットの特徴については，国土交通省総合政策局による分類[10]を参考に，「自然環境」,「歴史文化」,「都市」,「休憩」,「エンタメ」の 5 種類とする．各特徴について以下の 6 段階で回答する．

・5: 絶対行きたい
・4: 行きたい
・3: できれば行きたい
・2: できれば避けたい
・1: 絶対行きたくない
・0: どちらでもよい．

知名度は,「とても有名な所が良い」,「少し人気がある所が良い」,「あまり知られていない所が良い」と「どれでも良い」の 4 種類の選択肢から選択する．コストは,「無料」,「安い方が良い」(1-1000 円),「普通」(1001-3000 円),「高い方が良い」(3001 円以上) と「どれでも良い」から選択する．滞在時間は「30 分」,「1 時間」,「2 時間」,「半日」,「1 日」から選択する，ただし，滞在時間はグループメンバー全員で行動するため全員同じ選択をしてもらう．

(2)は，(1)で入力された各メンバーの回答を分析し，メンバーごとに 5 件の観光スポットを推薦アイテムとして選択した後，それらを統合してグループに対する推薦スポットリストを生成する．各メンバーに対する推薦スポットの決定手順として，最初に知名度，コストと滞在時間を満足するすべての観光スポットをデータベースから抽出する．知名度，コストに関して「どれでもよい」と回答した場合には全ての観光スポットがその条件を満たすとみなす．

データベースに登録されている観光スポットは，前述の観光スポットの 5 種類の特徴について，該当する度合い（関連度）を 1～5 の 5 段階のスコアで持っている．観光スポットXが持つ「自然環境」,「歴史文化」,「都市」,「休憩」,「エンタメ」の関連度をそれぞれ$x_1,\dots,x_5$とし，メンバーYの回答をそれぞれ$y_1,\dots,y_5$とすると，X,Yの非類似度$\mathrm{d}(X,Y)$は以下の式で定義される．

$$\mathrm{d}(X,Y)=\sqrt{\sum_{y_i\neq 0}(x_i-y_i)^2} \quad (1)$$

$d(X,Y)$の値が小さい順から観光スポットを 5 件選び, メンバーYに対する推薦スポット集合とする. 全メンバーについて求めた推薦スポット集合の和集合を求め, これをグループに対する推薦スポットリストとする.

(3) は, 推薦スポットリストを確認する順番を, 2.2 節で述べた MBTI 分類法に基づくアンケートを行い決定する. アンケートの内容及び順番の決定方法については 3.1 節で述べる.

(4) は, (3) で決めた順番に基づき, 一人ずつ順番に推薦スポットリストを確認・評価する. 具体的には,「とても気になる」,「気になる」,「普通」と「絶対行きたくない」の 4 段階で推薦スポットリスト内の各観光スポットを評価してもらう. 各メンバーに提示する推薦スポットリストは, それまでに確認したメンバーの評価に基づき, 以下のように更新される.

・一人以上が絶対行きたくないと回答した観光スポットを除外する.

・「とても気になる」の評価 1 件を+2,「気になる」を+1 として各観光スポットのスコアを求め, スコアの降順に並び替える.

リストを確認する際, 各観光スポットについて, 自分より前に確認したメンバーの評価を確認することができる. 最後のメンバーの評価に基づき更新されたリストが, グループに対する推薦スポットリストとなる.

3.1 MBTI による推薦スポットリスト確認順序の決定

提案手法では, 推薦スポットリストを確認する順番が重要となる. 本節では, 2.2 節で述べた MBTI 分類法を利用したアンケート結果に基づいて決定する手法を提案する. 具体的には, ユーザの感覚・直感と思考・感情の 2 指標を利用して, メンバーを以下の 3 種類に分類する : NT（分析家）, NF（外交官）, S（番人又は探検家）. 確認順序としては, NT タイプの人を最初, NF タイプの人を次, S タイプの人を最後にする. NT タイプの人は合理的で公平な思考を持っているため, 偏った判断はしないことが期待できる. NF タイプの人は共感力が強く協力的なため, 自分より前に確認したメンバーの評価を尊重しつつ, 後に確認するメンバーにも配慮して評価することが期待できる. S タイプの人は現実的な思考に基づくため, 他のメンバーが選択した結果として提示されるスポットの中から選択することに抵抗が少ないと考える.

上述の 3 タイプにメンバーを分類するため, 感覚・直感と思考・感情に関する設問だけを利用してアンケートを行うことを考える. そこで, MBTI に関する性格診断テストなどを参考に, 以下の 9 問を用意した.

感覚・直感に関する問題 :

1. 旅行に行くときはかなり計画を練る方である.
2. 他人の感情に共感することは難しいと感じることがよくある.
3. 討論において, 人の感受性よりも真実の方がより重要である.
4. 自分の行動が他人に及ぼす影響について心配することは, めったにない.
5. 他人に自分の行動についてあれこれ言わせない.

思考・感情に関する問題 :

6. 感情を支配するというより, 感情に支配される方である.
7. 詳細な計画を立てるのに時間を費やすよりも, どちらかというと即興で物事を実行する.
8. チームワークという点では, 協力的であるということの方が, 正しいということより重要である.
9. 皆の意見は, 事実の裏付けがあるかどうかに限らず尊重されるべきであると考える.

各設問に対し,「完全に同意する」,「同意する」,「大体同意する」,「わからない」,「あまり同意しない」,「同意しない」と「完全に同意しない」の中から選択して回答してもらう.

グループメンバーの負担を考えると, 設問は少ない方が好ましいと考える, そこで, 各設問の有効性, 必要性について検討するために予備実験を行った. 工学系大学生と大学院生 20 名に回答してもらった結果を表 1 に示す. 表では, 各設問について, 同意（「完全に同意」～「大体同意」）, わからない, 同意しない（「あまり同意しない」～「完全に同意しない」）にそれぞれ分類される回答の割合を示している.

表 1　予備実験結果

設問	同意する	わからない	同意しない
問題1	0.65	0.00	0.35
問題2	0.15	0.10	0.75
問題3	0.55	0.10	0.35
問題4	0.10	0.00	0.90
問題5	0.45	0.00	0.55
問題6	0.45	0.00	0.55
問題7	0.35	0.05	0.60
問題8	0.40	0.10	0.50
問題9	0.45	0.00	0.55

表より, 問題 2, 4 の回答は「同意しない」に偏り, 分類に有効ではないとられるため, 感覚・直感に関

する設問は問題 1，3，5 を選択する．問題 6，7，8，9 は回答の割合は類似している，しかし，問題間で回答の相関を調べたところ，問題 9 と問題 6,7 が負の相関となったため，思考・感情に関する設問は問題 6，7，8 を選択する．

各問題に対する回答から性格を分類するため，NS 値と TF 値を定義する．NS 値は，問題 1，3，5 に対する回答が「同意する」の場合は＋1,「わからない」場合は 0,「同意しない」場合は－1 を加算して求める．NS 値が正の場合は直感的，負の場合は感覚的であることを意味する．同様に，問題 6，7，8 に対する回答から TF 値を求める．「同意する」の場合は－1,「わからない」場合は 0,「同意しない」場合は＋1 を加算する．正の値は思考的，負の場合は感情的であることを意味する．

推薦スポットリストの確認順序は，前述の通り NT タイプの人を最初，NF タイプの人を次，S タイプの人を最後にする，具体的に以下の手順で決定する：
(1) NS 値，TF 値が共に正の人を抽出し，NS 値の降順に確認する．NS 値が同点だった場合は，TF 値の降順に確認する．
(2) 残ったメンバーから，NS 値が正のメンバーを抽出し，NS 値の降順で確認する．NS 値が同点だった場合は，TF 値の降順で確認する．
(3) 残りのメンバーについて，NS 値の降順に確認する，NS 値が同点だった場合は，TF 値の降順に確認する．

4. 評価実験

4.1 実験概要

提案手法の有効性を検証するため，工学系大学生と大学院生 20 名に提案手法を実装した推薦システムを使用してもらい，評価を行った．3 人，5 人，10 人のグループを構成してもらい，個人に対する推薦スポットリスト，グループに対する推薦スポットリストそれぞれについて満足度を評価してもらった．また，推薦リストを確認する順番をランダムにする対照実験も行い，結果を比較することで提案システムの有用性を検証する．実験協力者が行う手順は以下のとおりである．

Step 1 : メンバー各自で性格判断のための設問に回答する．
Step 2 : 一人ずつ順番に①，②を行う．
①行きたい観光スポットの特徴，知名度，コストと滞在時間を選択する．
②推薦スポットリストを確認し，満足度を評価する．
Step 3 : 指定された順序で一人ずつ順番に以下を行う．表示される推薦スポットリスト（図 1）の各観光スポットについて，「とても気になる」,「気になる」,「普通」,「絶対行きたくない」の 4 段階で評価する．
Step 4 : グループに対する推薦スポットリストを各メンバーで確認し，満足度を評価する．

Step2 の②は，メンバー個別に推薦した場合の評価を得るために行う．また，Step3 の実行順序は，提案手法により決定した順番，およびランダムに決定した順番のいずれかである．また，どのように順番を決定したかは説明せずに実験を行った．

図 1　観光スポット確認・評価画面

4.2 実験結果

実験結果を表 2 に示す．表において，「グループ人数」はグループを構成するメンバーの人数，「グループ数」は実験を行ったグループ数である．「個人」は Step の②で行った評価の平均値，「グループ」は Step4 で行った評価の平均値である．満足度に関する評価は 5 段階で行い，5 が最も高評価を意味する．

表より，3 人でグループを構成した場合，提案手法，ランダム順序どちらの場合も，個人に対する推薦よりもグループに対する推薦の方が評価が高くなる傾向にある．また，ランダム順序の方が評価が高い傾向にある．

表 2　実験結果

グループ人数	グループ数	提案手法		ランダム	
		個人	グループ	個人	グループ
3	5	2. 93	3. 47	3. 47	3. 60
5	4	3. 55	3. 95	3. 45	4. 15
10	2	3. 30	3. 60	3. 05	2. 70

5 人でグループを構成した場合は 3 人でグループを構成した場合と同様に，グループに対する推薦の方が評価が高い傾向にある．一方，提案手法とランダム順序の差は小さいといえる．

10 人でグループを構成した場合，提案手法ではグループに対する推薦の方が評価が高くなっているのに対し，ランダム順序では低下している．また，提案手法の方がランダム順序よりも評価が高い傾向に

ある．これらの結果より，グループの人数が多い場合に提案手法が有効に機能しているといえる．一般に，人数が多い程合意が難しくなると考えられるため，確認順序の違いが結果に明確に表れたと考える．

グループの人数によらない傾向として，10 人グループがランダム順序で確認した場合を除き，個人に対する推薦リストよりもグループに対する推薦リストの方が高評価されている．この理由に関しては，確認順序が早いメンバーの場合には，個人に対するよりも多数の観光スポットが推薦されていたと回答があった．また，最後の方に確認した場合には，他のメンバーが選んだ結果として良いスポットが残っていたという回答があった．

10 人グループにおける提案手法の有効性について考察するため，グループに対する推薦スポットリストへの評価と，確認順序の関係を図 2 に示す．ランダム順序の場合，最初と最後に確認したメンバーの評価が低い．原因として，最初に確認したメンバーは「スポットの数が多く，また下の方に気になっているスポットが多かったので大変だった」と回答していた．また，最後に確認したメンバーは自分の検索結果があまり反映されておらず，行きたいスポットが少なかったと回答していた．

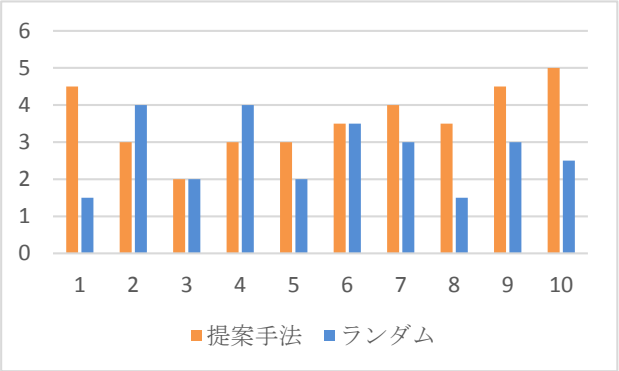

図 2　10 人グループにおける確認順序と評価の関係

確認順序と，Step3 で提示される観光スポット数の関係について，10 人グループの場合を図 3 に示す．前述のメンバーの回答にあったように，確認順序が 3 番目程度までは確認しなくてはならないスポット数が多いのに対し，最後の数名にはほとんどスポットが提示されていないことがわかる．

一人でも「絶対に行きたくない」と回答したスポットは表示されなくなるため，自分の希望を優先する傾向にある人が先に確認すると，後のメンバーの選択肢が減ってしまう．また，選択肢が多い方が好ましいかどうかについても，性格により違いがある可能性がある．以上より，グループ人数が多い場合は確認順序が結果に大きく影響したと考える．

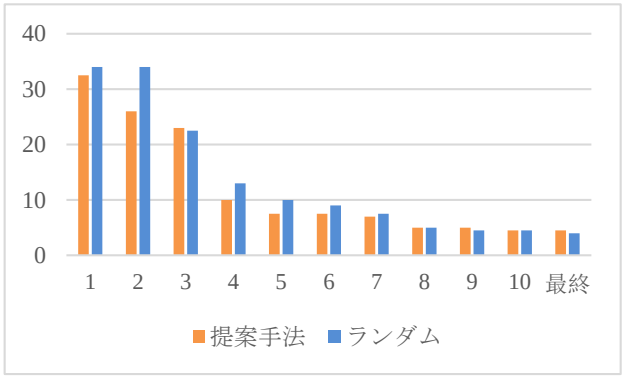

図 3　10 人グループにおける確認順位と平均確認スポット数の関係

表 3　10 人グループの推薦スポットリストの評価

順位	提案手法	ランダム
1	4.0/5.5	1.5/5.0
2	3.5/3.5	0.0/6.0

10 人グループにおいて，推薦スポットリスト上位のスポットに対する各メンバーの評価を表 3 に示す．表において，「とても気になる」と回答した平均人数を / の前に，「気になる」と回答した平均人数を / の後に示している．表より，提案手法の方が，多くのメンバーに支持されたスポットが推薦されていることがわかる．

5. まとめ

本稿では，グループを対象とした合議不要な観光スポット推薦手法を提案した．提案手法では，MBTI 分類法により推薦スポットリストを確認する順番を決定する．異なる人数のグループでユーザ実験を行った結果，グループの人数が多いときに提案手法の有効性を確認した．

今後は，より大人数のグループへの適用可能性を検討する．また，グループでの意思決定は国民性によって違う可能性があるため，国籍による提案手法の有効性の違いについて検討する価値もあると考える．推薦システムに関しては，観光スポットの特徴の追加やユーザによる評価方法などの改善を検討する．

参考文献

[1]　松本義之、藪内賢之: Web からの地域・観光情報収集

とその有用性の検討, 地域共創センター年報, Vol.7, pp1-17 (2014)

[2] I. B. Myers and P. B. Myers: Gifts Differing: Understanding Personality Type, Nicholas Brealey (1995)

[3] 鈴木綾子, 伊藤史子: 個人嗜好を考慮した訪問エリア選択支援システム-越後妻有大地の芸術祭における実証実験報告, 都市科学研究, No.4, pp.53-60 (2012)

[4] 嶋田和孝, 上原尚, 遠藤勉: 集合知に基づく観光地推薦システムの構築, 観光と情報, Vol. 10, No. 1, pp.113-124 (2014)

[5] 奥薗基, 牟田将史, 平野廣美, 益子宗, 星野准一: 複数人の嗜好分析による観光地推薦システムの提案, WISS2014, pp.147-148 (2014)

[6] C. G. Jung (G. Adler, R.F.C. Hull eds.): Collected Works of C.G. Jung, Volume 6: Psychological Types, Princeton University Press (1976)

[7] 生島淳: ラグビー日本代表ヘッドコーチ エディー・ジョーンズとの対話, 文藝春秋 (2015)

[8] R. M. Felder, G. N. Felder, E. J. Dietz: The Effects of Personality Type on Engineering Student Performance and Attitudes. Journal of Engineering Education, Vol. 91, No.1, pp.3-17 (2002)

[9] D. M. Walker, J. Fitz-Enz, J. A. Landry, J. F. Luebke, S. Bumpass, C. Filling, M. Gerber, E. Leinfuss, G. Buckalew, D. Carrington, B. Kutik, J. Monaghan: HR Director the Arthur Andersen Guide to Human Capital, Arthur Andersen (2000)

[10] 国土交通省総合政策局: 魅力ある観光地域づくりの秘訣 ～地域の取組をつなぎ・効果を高めるヒント集～ (2008)

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第19回)
SIG-AM-19-04

データの統合化と視覚化によるデータ分析統合ツールPADOCの提案

Proposal for Data Analysis Tool PADOC by Data Integration and Visualization

中井 眞人[1*]　角田 善彦[1]　林 久志[1]　村越 英樹[1]
Masato NAKAI[1]　Yosihiko TSUNODA[1]　[1] Hisashi HAYASHI　Hideki MURAKOSHI[1]

[1] 産業技術大学院大学 産業技術研究科
[1] School of Industrial Technology, Advanced Institute of Industrial Technology

Abstract: In Analects, there is a saying "visiting old, learn new". This means to investigate old things, in order to obtain new knowledge and insights.The process of data analysis could be interpreted as an act to analyze the data of the past, discover new knowledge and insights mathematically and make good use of them toward better future. Until end of the 20th century data was very valuable and less reliable, but now the accumulation of data became remarkable due to the explosive spread of the Internet society in recent years. However, proper use of data has not yet been established. The reason for this is that since the data is accumulated according to the operation of each business, there is no standardized analytical method because the accumulation state of data varies. Therefore, it is necessary to edit and integrate data by processing on computer for analytical purpose. Furthermore, it is necessary to examine whether the edited data is appropriate for analysis. To do so, it is convenient to have a tool that analyzes data while visually showing and checking data by editing and examining data. This paper proposes a graphical analysis integration environment "PADOC" to facilitate data editing and data review.

1　はじめに

論語の「温故知新」は昔の事を調べて，そこから新しい知識や知見を得ること（大辞林）[1] であるが，データ分析は過去のデータを分析して，その知見を数理的に発見することともいえる．20 世紀の後半まではデータは分析目的に合わせて収集することが多く，データ量も少なく信頼性も低いため，結果を出しても検証が煩雑であった。しかし近年のネット社会の爆発的な広がりでデータの蓄積は著しいものになったが，未だにデータ分析の適切な方法が確立されていない．これはデータが個々の業務の運用に合わせて蓄積され，データが多種多様に存在する一方，分析に適したデータが直接見つかることは稀で，見つかっても分散されている場合が多いからである．現在はデータ量は多くなったが，分析したいデータを作成するにはデータを適切に加工し統合することが必要になっている．このデータの加工と統合は前処理と云われ，一般的には全工程の 7 割が前処理に費やすと云われている．

*連絡先：産業技術大学院大学 産業技術研究科 創造技術学科
〒 140-0011 東京都品川区東大井１丁目１０－４０
E-mail:b1617mn@aiit.ac.jp

しかし近年急速に普及した無償の分析ツール R や Python は数理モデル構築に重点を置いており，データ編集が容易でないことが多い．そのため R や Python の前処理に特化した詳細な解説本「前処理大全」[2] が最近出版されている．

本稿はデータ編集とデータ分析を容易にするためのグラフィカルな分析統合環境 PADOC[1](Process Analysis by Data Oriented Composition) を提案する．これはデータ編集にはコマンド環境を提供し分析には分析過程をイメージし易い様にグラフィカル環境の両方を提供している．図 1 の左図は提案ツールのデータ編集のコマンド記述の一部を選択して実行している例で，右図はグラフ上のアイコンを繋いでデータ分析過程を構築している例である．

本稿ではデータ分析の目的を全体像の把握，比較検討，仮説検証，知識発見に分ける．全体像の把握に要するデータ編集では「前処理大全」にある Python と提案ツールの記述との比較を行ないその簡潔性を示す．比較検討や仮説検証及び知識発見では提案ツールが十分な機能を提供しグラフィカルな環境が効果的であることを示す．

図 1: Left:command mode Right:graphical mode

2　先行データ分析ツール

データ分析ツールは古くから有償の SAS[3]，SPSS[4]，S-PLUS[5]，Matlab[6] があり，近年では無償の R や Python が広く使われる様になっている．SAS は 1976 年から統計パッケージとして販売され最も実績がある．その実績から米国の医薬系の申請では SAS での報告が求められてきた [7]．SAS はデータが貴重で計算機が貧弱な時代に出現したので，結果の検定は得意としているが，高性能な計算が必要なグラフィカルな表示や直近のアルゴリズムの実装が貧弱である．逆に貧弱な計算機でも稼動する様にデータはテーブル形式を対象にしており，データ加工は一行単位に同じプログラムを適応するだけなので手続きが明瞭で記述し易い．関数型の S-PLUS は現在の無償の R として発展した．Matlab は行列演算を得意とし豊富な科学計算ライブラリィを提供したが，無償の Python も殆ど同等の機能が提供される様になり差別化が困難になってきた．R は初めて無償で本格的な統計パッケージであり広範に広まったが，関数型の記述や大規模なデータを不得意としているので Python に比べて劣勢にある．

一方隆盛を見せている無償の Python は高速な計算機とメモリーを贅沢に使ってプログラム言語の宣言やメモリー管理を不要にしてロジックが組みやすい記述を提供している．また無償化によって最新の深層学習系のロジックや豊富なグラフィック表現が次々提供されて加速度的に発展している．しかしオブジェクト指向型プログラム言語なのでオブジェクトに依存した多様なメソッドを駆使しなければならず習熟が難しい．またビックデータの前処理に既存のＰＣを連結して分散処理する無償の Hadoop[8] がある．これはクラウドの様な刻々と大量に流入するデータの選別やデータの変換には威力があるがデータ整形だけで分析機能がない．

提案ツール PADOC はコマンドベースではデータ加工を SAS の様にレコード単位に記述する平易な表現を採用し，一方グラフィカルな視覚表示では分析過程や分析結果を評価し易い環境を提供している．

[1]windows 7 8 10 で稼動，Python インストール要

3　提案ツールの説明手順

近年ではデータ分析の精度を争うサイト Kaggle[9] が出現し，データ分析技術の向上に大きく寄与している．この分野での知見が広がるのは望ましいが，データ分析とは理論や技術を駆使して分析精度を競うものと認識される傾向がある．しかしこれにはデータの前処理が抜けており，結果も精度を競うだけで実務上大事なモデルの頑健性や結果への説明力が抜けている．

実務用にデータ分析を定義したものとして総務省の「高度 ICT 利活用人材育成プログラム開発事業（実践偏)」[10] の資料がある．この資料ではデータ分析の用途を次のように分類している．本稿ではこの項番に沿って提案ツールの機能を説明する．

4. 全体像の把握
5. 比較検討
6. 仮説検証
7. 知識発見

4. 全体像の把握については，提案ツールのデータ編集の容易性を示し，5. 比較検証　6. 仮説検定　7. 知識発見では提案ツールの提供モデルとグラフィカルな表示環境が十分な機能を提供していることを示す．

4　全体像の把握

分散されたデータでは全体像を把握し難いので，一般的にデータを編集して統合する前処理を行う．しかし統合するとデータの定義は拠り所を失うので，データ定義はシステム運用に従って確認を行う必要がある．データの誤解釈は後続する分析を無駄にしてしまうので極力避けなければならない．

4.1　全体像の把握ツール

全体像の把握は各データ項目の充足状態や分布状態から分析に耐えられるか見る場合が多い．提案ツールでは分析対象項目とその他の項目との関係の強さ順に充足状態や分布も表示するツール [11] が提供されている．図 2 の例はローン破綻と関係が高い項目のランキング表示で，持ち家状態 (home)，ローン金額 (amount)，貸出し期間 (mon) が高い順になっていて，各項目の分布状態も示されている．持ち家状態 (home) の分布では賃貸や借家などの流動性が高い先のローン破綻率が高いことが示されている．

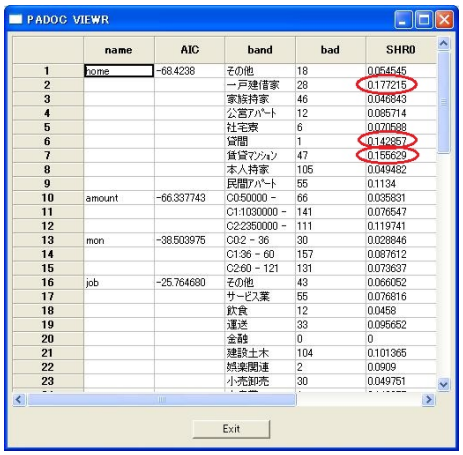

図 2: Relation ranking for loan collapse by AIC

4.2 データの前処理

データの前処理に関しては「前処理大全」の項目に沿って提案ツールの優位性を述べる．前述した「前処理大全」の前処理では次のセクションで記述されている．

(1) 抽出 (2) 集約 (3) 結合 (4) 分割 (5) 生成 (5) 展開

以下この手順に沿って提案ツールの前処理の優位性を述べる．

4.2.1 データ抽出（抽出）

データは業務運用のために蓄積されているので，分析用に必要な項目を抽出する必要がある．下記に示す様に欠損を除いて項目を抽出するには Python はメソッド関数を駆使するが，提案ルールは項目の列挙と抽出条件だけなので簡潔である．

- Python

```
#項目選択メソッドの使用
select_tb = bankr.loc \
[:['home','amount','job',]]
#欠損用の削除メソッドの使用
select_tb['amount'].dropna()
```

- 提案ツール PADOC

```
/* データ呼出し */
get bankr.csv@;
/* 項目選択 */
select home amount job;
/* 欠損レコードの削除 */
if(amount == ?) delrec;
```

4.2.2 サマリー処理 (集約)

一般にデータ分析では各レコードが独立であることが前提であるが，次の様な理由で独立を損なわれ結果に歪みが生じてしまう．そのためサマリー処理が必要である．

1. データが重複していると，重複が多いレコード寄りの結果となってしまう．

   例えば顧客別の明細に多寡がある場合，明細の多い顧客の特性が結果に反映されてしまう．この場合は顧客毎に明細をサマリーする必要がある．

2. 分析期間の長短によって結果が異なる．

   分析期間が異なると外的要因に晒されている期間が異なるので同じ条件のデータにならない．この様な項目は期間平均に直す必要がある．

以下は米国の業務種別（jobatnm)，人種別 (minoritynm) の給与 (salnow) をサマリーした例である．Python はメソッド関数を連結しているが，提案ツールでは，sumup コマンドを使って簡潔な表現でサマリー処理ができる．

- Python

```
reult = bankr.groupby \
(['jobcat','minority']) \
['salnow'].sum().reset_idex()
```

- 提案ツール PADOC

```
/* データ呼出し */
get bankr.csv@;
/* jobcat(職種) と minority(人種) 別にサマリー */
sumup  salnow by jobcat minority
```

4.2.3 データ結合 (結合)

データを統合するには分散データのＩＤを介して連結する必要がある．一般的には分散データ上で不要な項目を削除してから統合するのが普通である．この様な統合ではデータテーブル間の関係が容易にイメージできる必要がある．提案ツールではこの目的のためコマンドベースと図 3 の様なグラフィカル環境を設けている．図 3 は各分散データから項目を抽出して逐次合成する過程をアイコンで連結して実現している．一般に Python や「R」はオブジェクト指向のメソッド関数

を使った複雑な記述になり，全体的な統合過程をイメージするのが難しい．

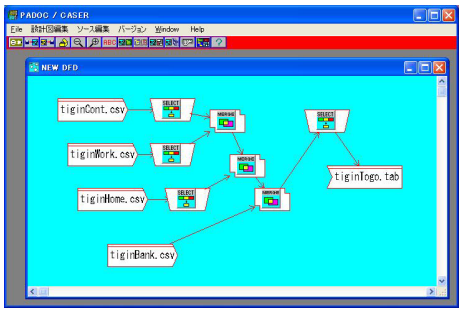

図 3: Data flow graph for Integration

4.2.4 データの分割（分割）

分散データを統合したデータをマスターデータと云うが，一般には以下の理由でデータ分割する場合が多い．

1. 分割による欠損の排除

   一般に統合すると業務上の理由で多くの項目で欠損が存在する．例えば株式会社であれば財務データは存在するが，個人営業会社では得られない場合が多い．この場合は株式会社と個人営業会社に分割して分析すると欠損を回避できる．

2. 分析の過学習を回避

   データ偏在を避けるためデータをランダムに分割し混合する交差検証でモデルの安定性を図る．

ランダムにデータを分割する場合は，提案ツールでは明示的に乱数の範囲を指定する

- Python

```
row_no = list(range(len(bankr)))
#4 分割指定
k_fold = KFold(n_splits=4,shuffle=True)
#4 分割
for train_cv_no in k_fold.split(row_no) :
  bank = train_data.iloc[train_cv_no,:]
```

- 提案ツール PADOC

```
get bankr.csv@;    /* データ呼出し */
rnd = random;      /* 一様乱数付与 */
/* 4 分割 */
     if(rnd <= 1/4) outrec bank1;
     else if(rnd <= 2/4) outrec bank2;
     else if(rnd <= 3/4) outrec bank3;
     else                outrec bank4;
```

4.3 データ加工（生成 展開)

データ加工の目的は分析に合う様に項目を作り出すことである．業務運用上蓄積されているデータだけでは，分析目的に合う項目が存在するとは限らない．分析用のデータを新たに生成する場合がある．

1. 市場データ等の公開若しくは有償で入手できるもの

   倒産の予測 [12] では日銀の市場データ [13] と自社の倒産推移とで図 4 の様に金利が倒産の推移に一年先行している事を使って 1 年後の倒産予測をしている．

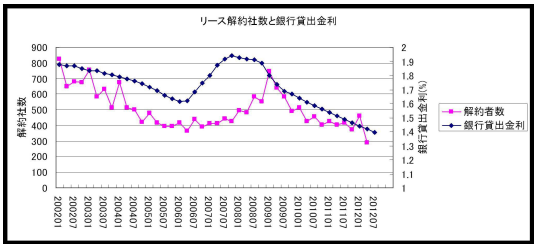

図 4: Relation of bankruptcy and interest rate

2. 自己資本比率の様に項目間で計算できるもの

   自己資本比率 = (総資本 - 負債)/総資本

3. 外れ値の補正やデータ値の偏在を避けるため対数化や正規化する

   一般に値段等の正の値を持つものは対数化すると正規分布になることが知られている．下記は対数化した例であるが，Python はオブジェクト志向型の言語なので，オブジェクト毎のメソッド関数を駆使する必要がある．一方提案ツールは平易な表現で全レコード対数化できる．

   - Python

```
reserve_tb['total_price_log'] = \
reserve_tb['total_price']. \
apply(lambda x:np.log(x/1000+1))
```

   - 提案ツール PADOC

```
get reserve_tb; /* データ呼出し */
togal_price_log=log(total_price/1000+1);
```

5　比較検討

本節以降の比較検討，仮説検討，知識発見について提案ツールの分析モデルは，グラフィカルな表現によって十分な性能を提供していることを示す．

一般に比較検討は区分による相違を見ることが多い．図 5 の右図のろうそく図は例は職種別の給与の分布を示し，一般職（左 3 業種）と資格職（右 4 業種) とで大きな相違が見られる．

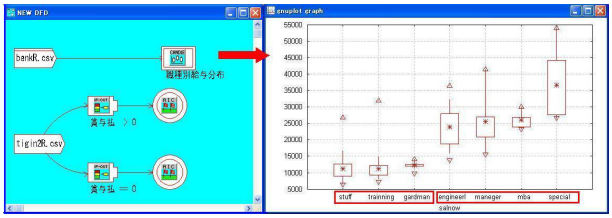

図 5: Candle chart for salay by job category

上図 5 の左図の分割プロセス図に示す様にローン返済で賞与払いの有無でデータ分割して，ローン破綻に関してランキング表示すると全項目について一度に比較することができる。図 6 の左図は賞与払い先で一定の賞与が見込まれる従業員のデータと見られ，右図は賞与払いが難しい経営者のデータと考えられる．ローン破綻が一番強く関係するのは従業員は借入金額 (amount) の多寡で，経営者は住宅の所有状態 (home)（即ち居住流動性）と相違している．

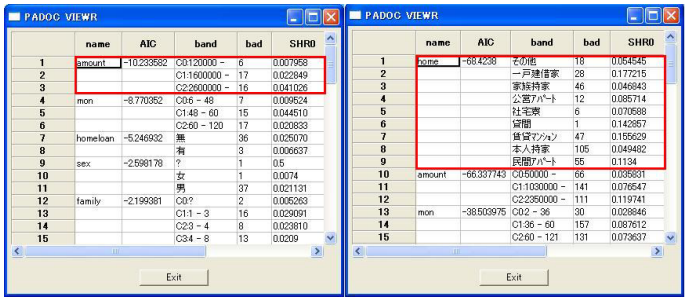

図 6: AIC ranking. Left:bonus Right:non bonus

6　仮説検証

仮説検証は，分析対象項目を他の項目で予測する仮説モデルの信頼性を検証する場合が多い．このモデルは分析対象データに合う様に予測するので一般には教師付モデルと云う．検証は予測値と分析対象の実測値の差を示す精度指標で行う．

6.1　教師付モデル

1. 分析対象が 2 値ならロジステック回帰モデルを使う [15]

   しかし図 7 の左図の様に全く線形的な分離が難い場合は SVM[14] が適切である．提案ツール SVM は右図の様にデータを高次元に写像してから線形分離している．

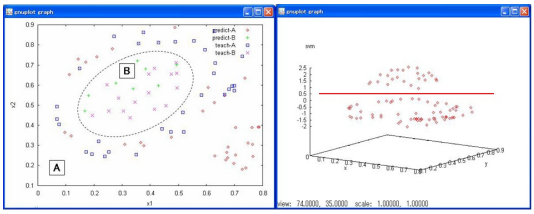

図 7: Impossible classification zone by linear

2. 分析対象が実数値なら重回帰モデル

   一般に重回帰は決定係数で検証する．図 8 の左図は 3D 図での回帰結果である．3D では回帰値は網掛けの平面になり、この平面上に殆どの点が載っていることがわかる．

3. 分析対象が時間経で劣化するならハザードモデル [16]

   図 8 の右図の例は格付別の会社の生存件数の経過予測したもので，下位格付ほど生存件数の降下が大きいことを示すことができている．

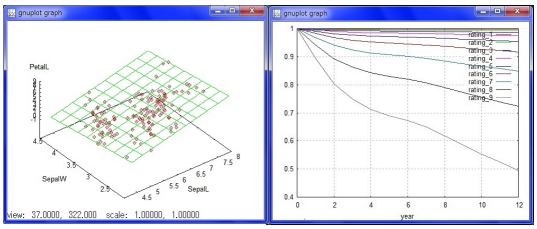

図 8:　Left 3D regression Right:cox hazard for bankruptcy

4. 時系列の推移予測で定常性があるなら ARIMA モデル，非定常性ならカルマンフィルターが使われる [25]

   提案ツールでは図 9 の左図の様にアイコンを繋ぐことでデータ加工過程を編集することができる．この図は株価をトレンド成分とフーリエで低周波を除去して，定常波にしてから ARIMA モデルで予測するプロセスを示したものである．右図の赤線部分は予測を示している．

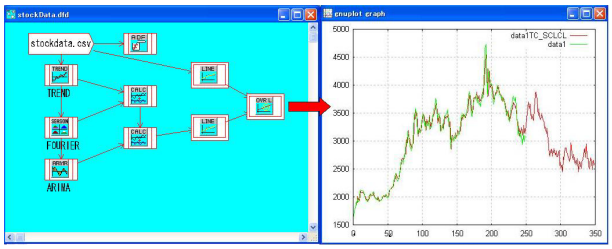

図 9: Left:analysis flow Right:prediction by ARIMA

6.2 精度と頑健性の問題

この様な教師付モデルは，分析対象と関係が強い項目（特徴量）の重み付線形和で表現されているので次の 3 つの観点で精度の検証が大切である．

1. 精度と頑健性はトレードオフの関係になっている．

   - 分析対象と関係の低い項目を追加しても精度は向上する．
     これはノイズでモデルを説明する過学習状態となっている．
   - レコードに重みを付与するブースティング法
     これは誤判別率に比して重みを付与して改善する方法である．

   何れも過学習すると試験用のデータでは精度は劣化する．精度と頑健性はトレードオフの関係があり，一般的には精度を高めると頑健性は劣化する場合が多い．

   分析対象が 2 値の場合では提案ツールは図 10 の左図の様に Boosting[17] を繰返して精度を上げることができ，右図の AR 曲線 [2] の様に Boosting による精度（緑線) と試験データでの精度 (赤線) を比較して，無理に精度を向上させていないか確認することができる．

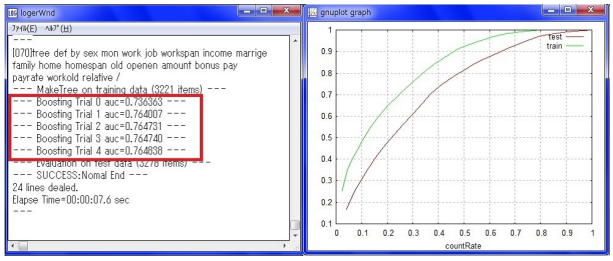

図 10: Right:boosting process Left:comparison of AR curves

[2] AR 曲線は横軸に事象の生起確率の高い順に並べ，縦軸に実際事象が起きた件数の累計をプロットして繋いだ線である．曲線の膨らみが大きい方が精度が良い．

分析対象が実数値の重回帰の場合は提案ツールでは正則化項を入れた Lasso モデルで過学習防止している．

2. 複数の分析手法を比較して優良モデルを選択する

   図 11 の左図は提案ツールでの判別木 (緑線) とロジット回帰 (赤線) の結果をグラフで重ね合わせて比較する過程を示した図である．右図は AR 曲線による精度の比較結果である．

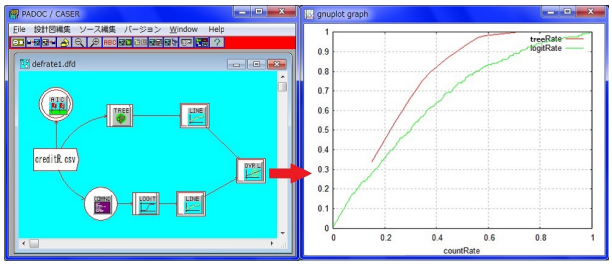

図 11: Right:process of Tree and Logit Left:comparision of AR curves

3. データが豊富にあればモデルの適合性を計るＡＩＣ [11] やＢＩＣ [19] を使わない．

   データが貴重な時代はモデルの適合指標として解釈が難しいＢＩＣやＡＩＣ等が使われたが，データが豊富な場合はデータを学習用と試験用に分割して図 10 の右図の様に精度比較をした方が合理的である．特に時間経過での頑健性を期待するなら，データを時間経過別にデータ分割して経年変化にも頑健性があるか検証すべきである．

7 知識発見

前節の仮説検証が目標データがある教師付モデルであれば，知識発見は教師データがなく一般的にはデータから隠れた要因を推定するモデルが主流である．知識発見では，隠れ変数推定ベイズモデル，グラフィカルモデル，最適計画問題がある．教師データが無いので結果の検証は難しい．結果は主に図やグラフで視覚的に説明される場合が多い．グラフィカルモデルはグラフ上でデータ項目間の最適な関係を示すものである．グラフに依らない最適問題として最適計画法や強化学習がある．提案ツールでは次の様なモデルを提供している．

1. 隠れ変数推定モデル

   1.1. MCMC(markov chain monte calro) [18]

隠れ変数をマルコフ連鎖でサンプリングして最大尤度を持つ値を探索する

1.2. EM 法 (expectation-maximization) [19]

隠れ変数の期待値で尤度を最大化する値を推定する

1.3. 変分ベイズ法 [19]

隠れ変数で仮定した変分の下界を最大化して値を推定する．

図 12 の例は米国の間欠泉 (Old Faithfull) の噴射間隔と噴射高のプロットしたものである．左図は EM 法　右図は変分ベイズ法で各々の集団の所属率を隠れ変数として推定している．

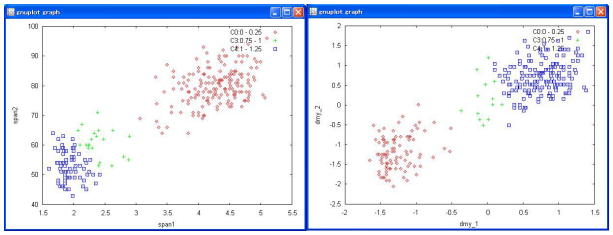

図 12: Left:EM Right:Variational Bayes for OldFaithfull

1.4. データの近傍状態を推定する樹系図

図 13 の左図の例は 10 人の成績がどの様に類似しているか示す樹系図である．赤で囲んだ Mia のみ離れているが右図の 3 D図でも外れ位置にあることがわかる．

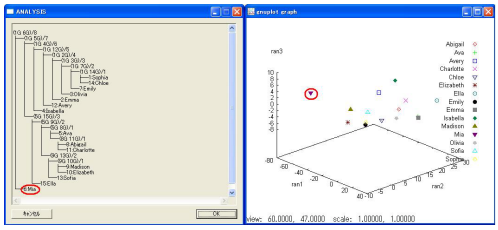

図 13: Left:dendrogram for students Right:3D plot by 3 subject score

1.5. 言語の類似状態 (トピック) を推定する LDA[20]

1.6. 時系列のパターンを推定する隠れマルコフモデル [21] 図 14 の左図は富士通の株価推移で，この隠れマルコフは 3 種の隠れモードを仮定している．右図の結果では一旦或るモードになると別のモードに移り難いことを示している．

図 14: Left:stock time series Right:Hidden Marcov mode

1.7. 異常値を検知する One Class SVM [22]

1.8. 推薦モデル GroupLense

図 15 の棒は 4 人の映画の評点を示す．左図では見ていない映画の評点は無い．右図は GroupLense が人と評点の類似性を見て評点を推定した結果である．見ていない映画の推定評点が高ければ推薦対象になる．

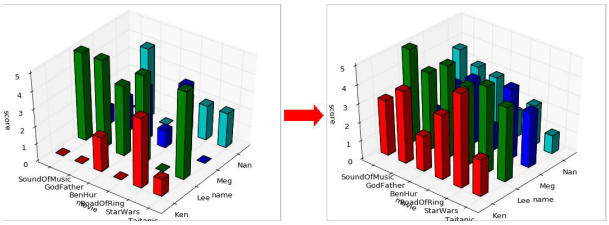

図 15: Left:exact score of movie Right:estimated score

2. グラフィカルモデル

提案ツールではデータから最適な関係をグラフで示すことができる．

2.1. ガウシアングラフィカルモデル

見かけ上の相関を排除した関係をグラフ表示する [23]　図 16 の左図の例は近代陸上 5 種競技データからの関係である．データ上は正の相関だが見かけ上の関係を排除すると負の相関が現れる．

2.2. 共分散構造 (SEM) モデル

項目間の関係を隠れた因子で説明するモデル [24] 図 16 の右図の例は社会活動のデータから社会的地位 (stage) を隠れ因子として説明した場合の関連の強さを表したグラフである．

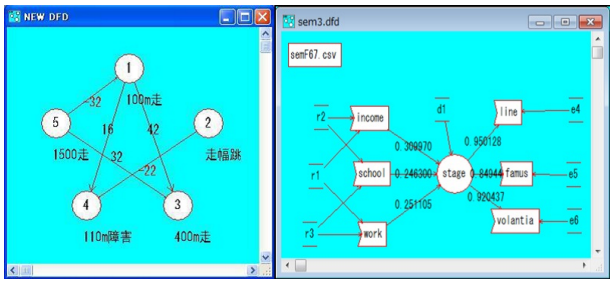

図 16: Left:5 athretic GGM Right: social status by SEM

2.3. ベイジアンネットワークモデル

データからベイジアンネットを自動生成する [26]

図 17 はローン債務者データからローン破綻 (def) を推定したベイジアンネットで．矢印に沿って確率伝播するので連結されたノードの条件に応じた確率が計算できる．

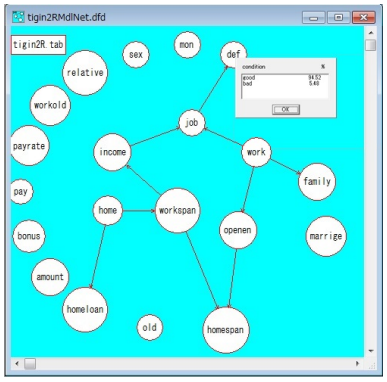

図 17: Baysian net for bank loan

2.4. 最短経路問題 (ダイクストラ法) [27]

図 18 の例は東京地下鉄の最短経路探索結果である．

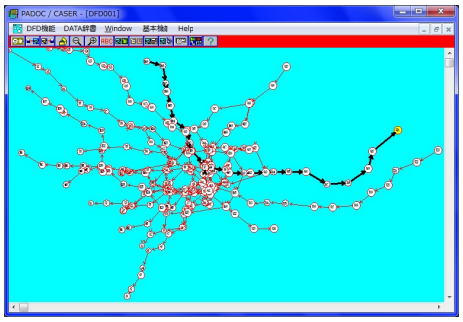

図 18: Optimize path in subway network of tokyo

2.5. 最大流量問題

図 19 の左図の例は各配管の容量制限内で配管網への最大流入量を示している．

2.6. 最大張木問題

図 19 の右図の例は全地点へ送電できる鉄塔の最短経路を示している．

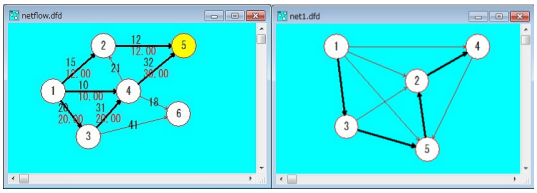

図 19: Left:max flow Right:max span tree

3. 最適化問題

提案ツールでは最適化問題として次の様な機能を提供している．

3.1. 線形計画問題

以下は実際の工場の人員配置問題を解いた例である．次の制約条件下で B 工場の最適な人員割当を求める．

＜ 制約条件 ＞

- A 工場は 8-17 時まで休まずに部品を作り順次 B 工場に搬送する
- B 工場は A 工場の部品の組立てと自工場でも部品製造と組立する
- 部品数以上に組立はできない
- 一日の生産数は，A 工場は 400 個 B 工場は 250 個
- 生産性は 1 人 1 時間当たり部品製造は 8.75 個 組立は 11 個
- B 工場では昼休みがあり 9 時台と 17 時台は 9 人 それ以外は 14 人出勤している

図 20 は B 工場の最適な人員割当の結果で，17 時台は 4 人省力化できることを示している．

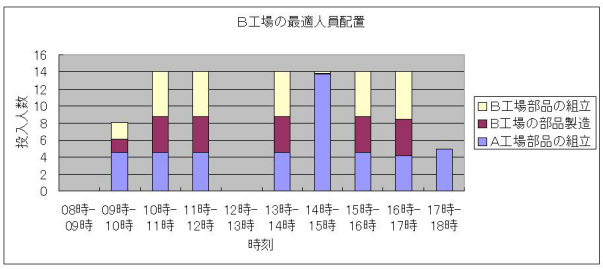

図 20: LienerPlan for personal allocation in factory

3.2. 整数計画問題 (分岐制限法) [28]

解が整数に限定される（ナップサック問題）

3.3. 非線形計画問題 [29]

図 21 の左図様な非線形問題を右図の様に条件式を設定すると最適問題を解くことができる．

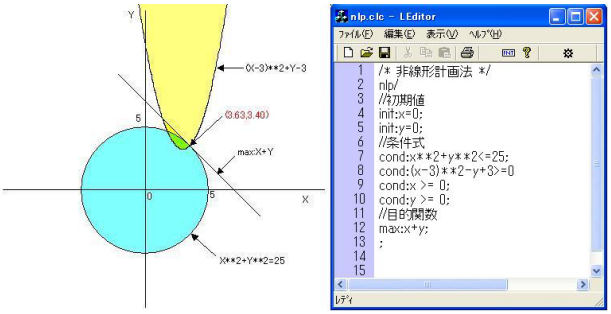

図 21: Left:promble Right:define nonleaner formula

3.4. 最適巡回路問題 (焼き鈍し法) [30]

図 22 は 3D のランダムな点を結ぶ最短順回路を解いた結果である．

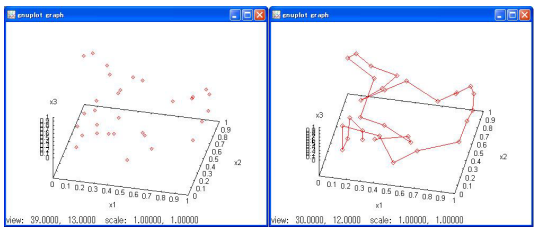

図 22: Left:random points Right:circuit path by anneal

4. 強化学習

最大報酬を得るために最適行動を学習するモデル [31]

図 23 の例は Sarsa 法による迷路探索の結果である [32]． この様に強化学習はゴールに達して初めて報酬が得られる場合でも解くことができる．

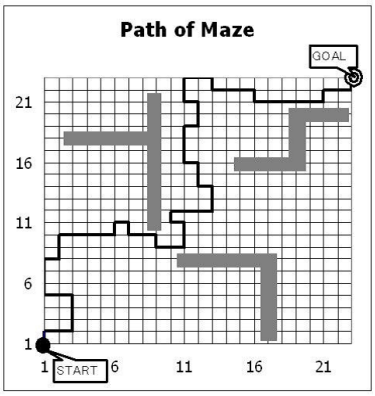

図 23: Reinforcement Learning for maze

8 考察

近年の機械学習やＡＩ関連の論文では殆どが Python で実装された実験報告になっている．これは Python がオブジェクト指向型の豊富な関数で複雑な数値計算ロジックを記述できるためと，直近の優良なモデルやライブラリィの公開が早く，これらを利用することで学術的な価値を高めるからである．逆に実用面としてデータの加工や編集の様な前処理をするには Python は記述が複雑すぎて一般的な技術者が利用するには負担が大きい．また現状では図 1 の右図で示す様なグラフィカル環境でモデルを構築する仕組みが Python には存在しない [3]．

実用的なデータ分析で要求されるのは高度な理論でなく，データ加工や編集を繰返して精度と頑健性があるモデルを構築し分り易く提示することである．提案ツール PADOC は平易な記述やグラフィカルな環境を提供し，試行の繰返しや評価を容易にしており実用的な面で優れていると考えられる．

9 まとめと今後

実用的なデータ分析として，全体像把握，仮説検定，比較検討，知識発見について，提案ツール PADOC を検討した．全体像把握については前処理大全 [2] の記述に沿ってデータ編集ツールとしての優位性を示し．比較検討，仮説検定，知識発見については提案ツールが提供するモデルがグラフィカルな環境によって性能をよく表す事を示した．

2017 年 DeepMind 社による棋譜学習しないでプロ級になる囲碁の学習モデル AlphaGoZero が発表された [33]．この論文の最後には「人類が数千年かけて習熟した囲碁の技を我々は数日で達成した」とある．機械学習の理論やモデルは革新的なものが日々研究され

ていると考えられる．この様なモデルをいち早く取り込める実用的なデータ分析環境として提供していきたいと考える．そのためPythonで作られたライブラリィを取り込むＡＰＩを公開して多くの人がこの環境をカスタマイズできる様にしたいと考えている．.

参考文献

[1] 孔子:『論語』為政第二 子曰．温故而知新．可以爲師矣

[2] 本橋智光, 前処理大全, 技術評論社 (2018)

[3] SAS Institute Japan 株式会社,SAS について https://www.sas.com/ja_jp/company-information.html

[4] IBM SPSS ソフトウェア https://www.ibm.com/analytics/jp/ja/technology/spss/

[5] 株式会社 NTT データ数理システム,S-PLUS for Windows https://www.msi.co.jp/splus/products/win/index.html

[6] MathWorks, 数学，グラフィックス，プログラミング https://jp.mathworks.com/products/matlab.html

[7] 木内貴弘,CDISC(Clinical Data Interchange Standards Consortium) 標準の概要 http://www.umin.ac.jp/indice/cdisc2009/01CDISC20091210pdf.pdf

[8] Tom White,Hadoop,Orailly(2013)

[9] データ分析競争サイト, https://www.kaggle.com/competitions

[10] 総務省,「高度 ICT 利活用人材育成プログラム開発事業（実践偏)」 データ解析手法とツール (2012) http://www.soumu.go.jp/main_sosiki/joho_tsusin/joho_jinzai/

[11] 坂本慶行 石黒真木夫 北川源四郎, 情報量統計学 § 6 分割表解析モデル, 共立出版 (1998)

[12] 長井章夫, マクロ経済指標を使った重回帰分析による倒産予測 SAS ユーザ会 2012

[13] 日本銀行, 時系列統計データ検索サイト, https://www.stat-search.boj.or.jp/

[14] John C. Platt,Fast Training of Support Vector Machines using Sequential Minimal Optimization

[15] 丹後俊朗, ロジステック回帰分析, 朝倉書店

[16] 中村剛,Cox 比例ハザードモデル, 朝倉書店

[17] Trevar Hastie,The Elements of Statical Learning,Springer(2008)

[18] 伊庭幸人, マルコフ連鎖モンテカルロ法とその周辺, 計算統計 2 , 岩波書店 (2005)

[19] Christopher M. Bishop:Pattern Recognition and Machine Learning § 11, Springer(2006)

[20] 持橋大地, 確率的トピックモデル http://www.ism.ac.jp/~daichi/lectures/H24-TopicModel/ISM-2012-TopicModels-daichi.pdf

[21] Mark Stamp:A Revealing Introduction to Hidden Markov Models(2012)

[22] 高畠泰斗 香田正人, 1 クラス SVM と近傍サポートによる領域判別 (2006)

[23] Hirono,Graphical Model http://www.mayomi.org/lecture01/BSJ2008/BSJ2008tutorial1.pdf

[24] 豊田秀樹, 共分散構造解析 [入門編], 朝倉書店

[25] 北川源四郎:時系列解析プログラミング—FORTRAN77 (岩波コンピュータサイエンス)

[26] 中井眞人, データからのベイジアンネットワーク構造推定, 人工知能研究会 2014

[27] 佐藤公男, グラフ理論入門, 日刊工業新聞社 (2004)

[28] 大山達雄, 最適化モデル分析, 日科技連出版社 (1993)

[29] 矢部博, 最適化とその応用, 数理工学社 (2006)

[30] Willam H. Press, Neumerical recipes Third edition § 10 (2011)

[31] Richard S. Sutton and Andrew G. Barto1,Reinforcement Learning:An Introduction

[32] 中井眞人,Sarsa(λ) 法による強化学習の汎用化 (2015) https://www.slideshare.net/MasatoNakai1/ros-63464994

[33] David Silver,Mastering the game of Go widthout human knowlege ,Nature(2017)

[3]但し深層学習ネットワーク構築用のグラフィカルツールとしてTensorBoard がある