複数文章からの表生成における生成 AI の利用と評価手法の比較

野田直哉 [1]　村田真樹 [1,2]
[1] 鳥取大学大学院 持続性社会創生科学研究科 工学専攻
[2] 鳥取大学工学部附属クロス情報科学研究センター
m23j4039b@edu.tottori-u.ac.jp
murata@tottori-u.ac.jp

概要

村田ら [1], 野田ら [2] は情報整理の一環として Web 上の複数の関連した文章から表を生成していた. これらの研究での表の評価として, 予め用意した正解の表との類似性で自動評価を行っていたが, 出力された表の構造が正解のものと異なる場合, 実際の表生成の性能を正しく評価できないと考えられた.

本稿の研究では, より信頼性のある評価を目指し, 人手で評価をした. 村田らと野田らの単語ベクトルと ChatGPT を使用した手法をそれぞれ評価したところ, 自動評価では単語ベクトルの手法のほうが性能が高かったが, 人手での評価では ChatGPT の手法のほうが高くなった. 人手評価の単語ベクトル, ChatGPT の手法の性能はそれぞれ 0.79, 0.90 であった.

また自動評価の問題である, 正解と出力の表の構造が異なる点について, 調査をしたところ, 実際に評価に影響を及ぼしていることがわかった.

1　はじめに

近年, Web 上のテキストデータの構造化や, 情報抽出に関する様々な手法 [3][4] が提案されている. 情報整理の一環として, 村田ら [1], 野田ら [2] は Web 上の関連した複数の文章から表を生成していた. これらの研究では, 図 1 の文章を, 表 1 のように重要情報を自動で抽出し表の形で整理していた. また, この表生成の応用として, 村田ら [5][6][7] の研究では, ChatGPT と表生成技術を用いて, 株価記事に対する分析を行っている.

村田ら [1], 野田ら [2] の研究では, 表の評価方法として, 正解をあらかじめ用意し正解との類似性で自動評価をしていた. しかし, この自動評価では正解との類似性を評価しているため, 出力された表の構造が正解のものと異なる場合, 実際の表の性能を反映していないと想定され, 自動評価の信頼性に疑問があった. 本稿では, 自動評価ではなく出力された表を人手で評価することによって, より信頼性のある評価をすることを目指した.

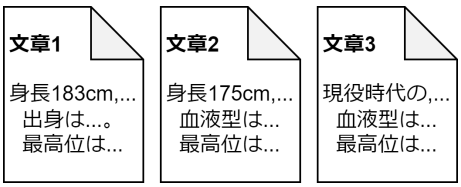

図 1　入力例 (力士)

表 1　出力例 (力士)

	身長	最高位
文章 1	身長 183cm, 体重 191kg	最高位は東十両 12 枚目
文章 2	身長 175cm, 体重 130kg, 血液型は O 型	最高位は西幕下 2 枚目
文章 3	現役時代の体格は身長 179cm, 体重 149kg, 血液型は AB 型	最高位は東幕下 4 枚目

2　関連研究

2.1　単語ベクトルによる表生成

村田ら [1] の研究では, 複数の文章から重要情報を抽出し表の形で整理して表示していた. 図 2 にあるように, 文の単語のベクトルの平均を文のベクトルとして, 文をクラスタリングし, 各クラスタの中から重要項目となりうるクラスタを選別しそのクラスタを列とする表を作成していた.

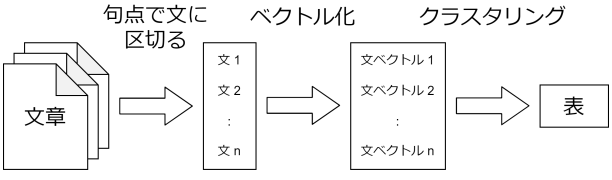

図 2　単語ベクトルによる表生成の手順

2.2 ChatGPT による表生成

野田ら [2] の研究では, GPT4 を用いて表を生成していた. 図 3 は, ChatGPT による表生成フローである. 表生成の対象となる文章を ChatGPT へ入力し, 表に用いる項目を出力させ, その項目を用いて, ChatGPT で各文に項目をラベル付けしている. そして, 文が属する文章を縦軸, 項目を横軸として, 各文を整理することで表を作成している.

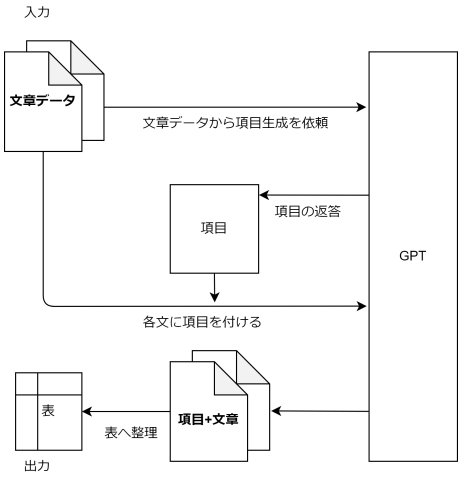

図 3 ChatGPT による表生成の手順

3 実験

単語ベクトルと ChatGPT による表生成の評価実験を次のような条件のもとで行った.

本稿の実験で正解として使用した表データは, 人手で作成した 15 個の表である. この表のもととなった記事は, 「エアコン」といったガジェット系の記事が 5 種類, 「地震」といったニュース記事が 5 種類, 「日本の城」といった wikipedia からの記事が 5 種類の計 15 種類である. 用いた ChatGPT のモデルは GPT4 である.

4 自動評価

4.1 自動評価の方法

自動評価では, あらかじめ人手で作成した正解と実験での出力を列ごとに比較し, 類似度を元に評価している. このとき, どの列を比較するかを判断するために, 正解と出力の列をマッチングさせる必要がある. 図 4 のように, 正解の列に最も類似している出力の列を正解の列にマッチングさせ, 列の比較をしている. また, 正解と出力の列の類似度として, 再現率 (Recall), 適合率 (Precision), F 値を列ごとに計算をし, 各列にある文数を重みとして求めた平均を表の再現率・適合率・F 値としている. この評価において再現率・適合率・F 値それぞれの意味は以下のようになる.

- 再現率: 取りこぼしなく列が作られているか
- 適合率: ある列に, 不適切な文が存在しないか
- F 値 : 再現率・適合率の調和平均

A を正解の列に含まれる文の集合, R を結果の列に含まれる文の集合とすると, これら 3 つの値は次のようになる.

$$\text{Recall} = \frac{|A \cap R|}{|A|} \tag{1}$$

$$\text{Precision} = \frac{|A \cap R|}{|R|} \tag{2}$$

$$\text{F} = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}} \tag{3}$$

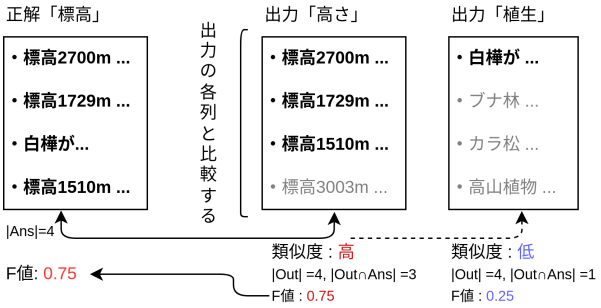

図 4 評価方法

4.2 問題点

同一の文章から作成された表であっても, 表 2 のように, 異なる項目が考えられる. このように項目が異なることで, 正解と出力の表全体の構造が大きく異なり, 現在の自動評価で使用している, 正解との類似性での評価では表生成の性能を正しく評価できない場合が想定される.

この構造の問題は実際の出力で確認されている. 表 3 の, 項目「発売日」に関して, 正解と一致しているが, 出力の項目「色/デザイン」と判定された文が正解の表では「掃除・手入れ」「カラー」「デザイン」の 3 つの項目で分類されている.

表 2　生成された項目の例 (エアコン記事)

正解の表での項目	出力された項目
価格	価格
発売日	発売日
省エネ	エネルギー効率/消費電力
ラインアップ	モデル名
カラー	色/デザイン
冷暖房性能	冷房能力（kW）
気流の制御	暖房能力（kW）
サイズ	適用畳数（冷房時）
掃除・手入れ	適用畳数（暖房時）
除菌・清浄	空気清浄機能
除湿	主な特徴
デザイン	
AI	
人物検知	
メーカー発表	
他の機器との連携	

表 3　項目と文の例

正解の項目	出力の項目	文
発売日	発売日	11 月 1 日より発売する。
発売日	発売日	3 月下旬より発売する。
発売日	発売日	10 月下旬より順次発売する。
掃除・手入れ	色/デザイン	吹き出し口周辺のルーバーやダストボックスなどは簡単に取り外し、水洗いが可能。
カラー	色/デザイン	ボディカラーは、パウダースノウとボルドーレッドの 2 色を用意した。
デザイン	色/デザイン	風の流れをイメージして曲面を基調とした「ウェーブデザイン」の前面パネルを採用。

また, 正解の表の作成者と異なる人が表を作成し, 人手による表が自動評価でどの程度の性能を得るかを確かめた. 表 4 は, 人手と単語ベクトル, ChatGPT のそれぞれの自動評価の結果である. 人が優れた表を作ると思われるが, この結果では人手の性能が高いとは言えず自動評価に信頼性があるとは言えない結果であった.

表 4　自動評価

手法	再現率	適合率	F 値
人手	0.88	0.48	0.61
単語ベクトル	0.73	0.63	0.67
ChatGPT	0.82	0.47	0.57

5　表の人手評価

5.1　人手評価の方法

自動評価ではあらかじめ作成した正解の表との類似性で評価をしており, 表の構造が正解と異なることで性能が大きく下がるなどの問題があり, 自動評価の信頼性に疑問があった.

本稿では, 自動評価ではなく出力された表を人手で評価することによって, より信頼性のある評価を実現することを目指した. 人手での評価は, 各列に含まれている文が適切であるかを人手で判断し, 適切と判断した文の割合を評価した. この評価を, 人手評価の適合率とする.

また, 本稿では, 評価のコストが大きいため自動評価の再現率に相当する評価を行っていない.

5.2　人手評価の結果

人手で適合率の評価をした結果を表 5 に示す. 適合率において, 自動評価では単語ベクトルの手法が最も高い性能であったが, 人手での評価では, ChatGPT のほうが高く, 異なる結果が得られた. 人手評価のほうが自動評価より信頼性があると考えると, ChatGPT による表生成のほうが性能が高い.

表 5　人手評価と自動評価

出力	人手 適合率	自動 適合率	自動 再現率	自動 F 値
人手	—	0.48	0.88	0.61
単語ベクトル	0.79	0.63	0.73	0.67
ChatGPT	0.90	0.47	0.82	0.57

5.3　出力例と評価

表 6, 7 は, 単語ベクトルと ChatGPT による手法で得られた結果と, その人手評価である.

表 6　単語ベクトルの出力と人手評価例

評価	項目「Bluetooth」from カメラの記事
○	Ｗｉ-Ｆｉ、ＮＦＣ、Ｂｌｕｅｔｏｏｔｈによる通信機能も搭載した。
×	接続端子はＭｉｃｒｏ ＵＳＢ端子を装備。
×	ＵＳＢ接続ケーブル、ハンドストラップが付属する。
評価	**項目「報告」from リコール案件の記事**
○	14 年 11 月以降、59 件の不具合が報告された。
○	ガソリン臭がするとの報告が 13 年以降に205 件あるが事故例はない。
○	これ以外の不具合は確認されていない。

表 7　ChatGPT の出力と人手評価例

評価	項目「山の名前」from 山の記事
○	燧ヶ岳（ひうちがたけ）は福島県にある火山。
○	幌尻岳（ぽろしりだけ）は、 ...
○	山名はアイヌ語で「大きな山」を意味する。
評価	**項目「メーカー名」from リコール案件の記事**
○	スズキは１３年６月、... 再リコールした。
○	東芝とパナソニックは ... リコールすると発表した。
×	問い合わせはフリーダイヤル０１２０・８７０１６３。

6 考察

6.1 人手評価と自動評価の結果

人手評価と自動評価の結果の具体例を表 8 に示す. 表 8 のエアコンの記事の例では, 自動評価と人手評価の結果が異なっている. 3 行目の文はデザインに書かれている文であるが, 自動評価の結果では正しくないと判断されている. このような評価となっている理由は, 出力の項目「色/デザイン」が, 正解では, 表 3 にあるように「カラー」「デザイン」という 2 つの項目に分割されており, 出力の項目「色/デザイン」を正解の項目「カラー」で評価されたためだと考えられる.

表 8 ChatGPT の出力の人手評価と自動評価の例

人手	自動	項目「色/デザイン」 from エアコンの記事
×	×	吹き出し口周辺のルーバーやダストボックス などは簡単に取り外し、水洗いが可能。
○	○	ボディカラーは、パウダースノウと ボルドーレッドの 2 色を用意した。
○	×	風の流れをイメージして曲面を基調とした 「ウェーブデザイン」の前面パネルを採用。
人手	自動	項目「接続端子」from スマートフォンの記事
○	×	このほか、接続端子は USB Type-C を採用。
×	×	背面には、指紋認証センサーを備える。
×	×	このほか、 ... 高性能指紋センサーを背面に用意。

6.2 自動評価の問題点の調査

自動評価に関する問題として, 4.2 節では, 正解との類似性を評価しているため, 出力された表の項目が正解のものと大きく異なる場合, 実際の表生成の性能を正しく評価できないと考えた.

この表の項目が大きく異なることの問題が評価に影響しているかを確かめるため, 問題が発生しないように, 正解と同じ項目を持つ表を出力させ, その表と通常の手法での表を比較した. 正解と同じ項目を持つ表は, ChatGPT に表の構造のヒントとなる正解での項目のリストを与えて生成した. その方法を本節では「ChatGPT ヒント有り」と呼ぶこととする. 「ChatGPT ヒント有り」と通常の手法での表に対してそれぞれ自動評価と人手評価をしたとき, 「ChatGPT ヒント有り」のほうが, 自動評価が人手評価に近づくと推測した.

「ChatGPT ヒント有り」と通常の ChatGPT と単語ベクトルの手法での表生成の適合率を, 15 個のデータからもとめて, その適合率 15 組で自動評価と人手評価の相関係数を求めた. 表 9 によると, 「ChatGPT ヒント有り」のほうが相関係数が高く, 自動評価と人手評価の結果が近い. 「ChatGPT ヒント有り」のほうが自動評価が人手評価に近づくという, 推測と矛盾のない結果が得られた.

この検証により, 出力された表の構造が正解のものと異なるという問題が, 実際の評価に影響を及ぼしていると考えられる.

表 9 自動評価と人手評価の相関

手法	人手 適合率	自動 適合率	相関係数
単語ベクトル	0.79	0.68	0.43
ChatGPT	0.90	0.57	0.49
ChatGPT ヒント有り	0.91	0.82	0.71

7 おわりに

村田ら [1], 野田ら [2] は同一テーマの複数の文章から項目ごとに情報を整理することで表の作成を行った. これらの研究では, 表を評価する際に, 人手で作成した正解の表との類似性をもとに自動評価をしていた. しかし, この自動評価では正解との類似性を評価しているため, 出力された表の項目が異なる場合実際の表生成の性能を正しく評価できないと想定され, 自動評価の信頼性に疑問があった.

本稿の研究では, 生成された表を人手で評価することにより, 信頼性の高い評価を目指した. 自動評価では, 「不適切な文が列に存在しないか」という指標として適合率を計算しており, 人手でこれを評価した.

人手で単語ベクトルと ChatGPT を使用した手法を評価したところ, 自動評価では単語ベクトルの手法のほうが性能が高かったが, 人手評価では, 単語ベクトル, ChatGPT の手法のので適合率はそれぞれ 0.79, 0.90 で, ChatGPT の手法のほうが性能が高かった. 人手評価のほうが自動評価より信頼性があると考えると, 単語ベクトルによる表生成よりも ChatGPT による表生成のほうが性能が高いと言える.

また, 自動評価が実際の表生成の性能を反映されない原因を, 正解と出力の項目が大きく異なることと想定し, 調査を行った. 調査では, ヒントを与えることで正解の表での項目と同じ項目を持つ表を生成し, その表では自動評価の傾向が人手評価の結果に近づくことを確認した. 正解と出力の項目の違いが自動評価に影響を及ぼしていると考えられる.

参考文献

[1] Masaki Murata, Kensuke Okazaki, and Qing Ma. Improved method for organizing information contained in multiple documents into a table. **Journal of Natural Language Processing**, Vol. 28, No. 3, pp. 802–823, 2021.

[2] 野田直哉, 村田真樹. ChatGPT を用いた複数文章からの表生成. 言語処理学会第 30 回年次大会発表論文集, pp. 2966–2970, 2024.

[3] Chia-Hui Chang, kayed Mohammed, Girgis Moheb, Ramzy, and F. Shaalan Khaled. A survey of web information extraction systems. **IEEE Transactions on Knowledge and Data Engineering**, Vol. 18, p. 1411–1428, 2006.

[4] Xiang Deng, Ahmed Hassan Awadallah, Christopher Meek, Oleksandr Polozov, Huan Sun, and Matthew Richardson. Structure-grounded pretraining for Text-to-SQL. **NAACL 2021**, pp. 1337–1350, 2021.

[5] Masaki Murata. Content analysis of items in newspaper data using table arrangement technology and ChatGPT for stock price prediction. In **Proceedings of The 22nd International Conference on Information & Knowledge Engineering on CSCE 2023**, pp. 1–8, 2023.

[6] Masaki Murata. Leveraging ChatGPT and table arrangement techniques in advanced newspaper content analysis for stock insights. **Big Data, Data Mining and Data Science: Algorithms, Infrastructures, Management and Security**, 2024.

[7] 村田真樹. ChatGPT と表整理技術を利用した株価に関わる新聞記事の分析. 言語処理学会第 30 回年次大会発表論文集, pp. 2852–2857, 2024.