iStar モデルの誤りを指摘する方法の提案

橋浦研究室　　122D015 泉 汰輝　　122D070 清水 優希

1　はじめに

システム開発における要求工程は，大きな手戻りや開発コストの増大を防ぐために重要である．その中でも要求分析は，要求の曖昧さや抜け漏れを指摘し，取り除くという点で要求工程の中核をなすものである．

要求分析手法である，図式言語を用いた iStar 2.0 [1](以下 iStar) は，関係者の利害や目的を明確化する手法として有効であることが知られている．しかし，モデリングの過程において，図 1,2 のような要求文とモデルで整合性が取れていないなどの誤りが発生してしまう．このような誤りを含んだ iStar モデルが下流工程で使用されることにより，機能の未実装や余分な機能の追加などの低品質なソフトウェアの開発に繋がる．

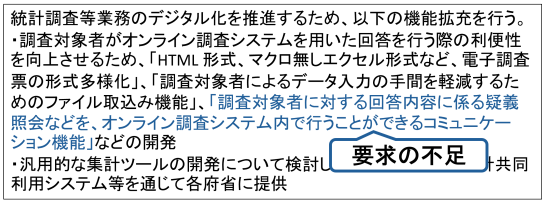

図 1: 要求文 [2]

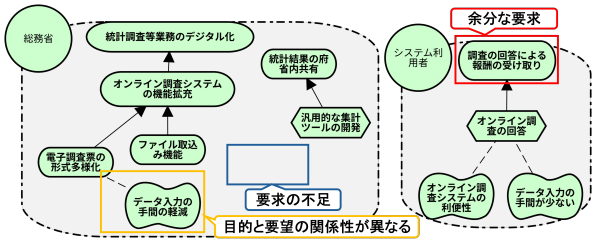

図 2: 誤りを含むモデル

2　関連研究

iStar モデルの誤りを見つける研究として，Hirabayashi *et al.* [3] が挙げられる．この研究では，モデルの品質低下の要因としての"bad smells"を定義し，モデルの構造的および意味的な観点から自動検出する手法を提案している．

この研究では，モデルの構造的な正しさを扱っているが，要求文との意味的な整合性までは扱っていない．

3　提案手法

本研究ではモデリング時に発生する誤りに対して，要求文の意味を考慮したうえで，機械的処理による自動発見を目指す．そのために，7 つに分類した誤りのカテゴリに該当する箇所を，機械学習や自然言語処理技術を用いて，要求文もしくはモデルから，誤り箇所を自動的に検出する．しかし，自動検出の結果は，確実な誤り箇所とは限らないため，誤り候補として扱い，潜在問題分析 [4](以下 PPA) を用いて，修正すべき確実な誤りであることをユーザーが検証する手法を提案する．

表 1: 誤りのカテゴリ

#	カテゴリ	要約
1	抽出した意図的要素の不足	要素の不足
2	無関係な意図的要素の作成	無関係な要素
3	関連付ける意図的要素の誤り	関連付けの誤り
4	意図的要素の種類の誤り	図形の誤り
5	関連の種類の誤り	線の誤り
6	抽出した意図的要素の重複	要素の重複
7	テキストの未入力	未入力

本研究で用いた誤りのカテゴリを表 1 に示す．カテゴリごとの検出方法として，カテゴリ 1，3，6，7 は，要求文とモデルを分析する．カテゴリ 1 と 3 は，要求文から抽出した動詞句とモデル要素の類似度を計算し，意図の不足や過剰な作成を検出する．カテゴリ 6 と 7 は，要求文中の記述や位置関係を分析し，要素間の関係の有無や種類の整合性を確認する．

また，カテゴリ 2，4，5 は，モデルでの整合性を担保するためにモデルのみの分析を行う．カテゴリ 2 は，要素間のベクトル類似度を用いて意味的な重複を検出する．カテゴリ 4 は，学習済みの Transformer モデルを用いて記述内容と要素種類の意味的な不一致を特定する．カテゴリ 5 は，デフォルトテキストのままの要素を検出する．

次に，自動検出による誤り候補に対し，PPA の分析を以下の項目に対して行う．

表 2: PPA の分析項目

#	カテゴリ
1	潜在的問題
2	該当理由
3	潜在的問題の発生頻度
4	潜在的問題の予防策
5	修正が必要か？(yes/no)

このリスク評価を通じて，ユーザは検出結果が修正すべき誤りであるか否かを判断する．

4　研究目的

本研究の目的は，iStar モデルの誤りを指摘できるようにすることである．そのために，提案する検出器の有効性を確認するための以下の Research Questions(以下 RQs) を設定する．

RQ1: 検出器による検出結果は正しいか
RQ2: 提案手法によって誤りの指摘が出来るか

5　評価

ツールの評価を行うために，日本工業大学の学部生 20 名，大学院生 3 名の計 23 人を被験者として，ツールを用いて誤り候補を検出するグループ (実験群) と手作業で誤り候補を検出するグループ (統制群) に無作為に割り振り実験を行った．また，被験者は，インターネットに公開されている要件定義書を用いてモデリングを行った後，誤り候補の検出と PPA の分析を通して，確実な誤りを指摘する．

RQ1 に回答するためには，ツール出力の妥当性を明らかにする必要がある．このため，著者らでツール出

力の妥当性の検証を手作業で行った後，κ 係数 [5] と Gwet's AC1 [6] を用いて作業の一致度合を評価する．

著者らが，実験群と統制群で作成されたモデルから，ツール出力されるべき誤りの要素 (誤った要素) とツール出力されない正解の要素 (正しい要素) を判断した．表 3 から，κ 係数は 0.5591，Gwet's AC1 では 0.8469 となり，著者らが行った手作業の一致度合は高いと評価できる．

表 3: 妥当性評価の結果

		著者 B		
	#	誤った要素	正しい要素	行-合計
著者 A	誤った要素	154	86	240
	正しい要素	98	1281	1379
	列-合計	252	1367	1619

表 3 の著者らの判断が一致している数のみを用いて，実験群と統制群におけるツール出力との一致数を表 4 に示す．

表 4: 2 群と妥当性評価の一致数

		妥当性評価結果		
	#	誤った要素 (割合)	正しい要素 (割合)	行-合計
実験群	検出数	50(0.7042)	321(0.4566)	371
	未検出数	21(0.2958)	382(0.5434)	403
	列-合計	71	703	774
統制群	検出数	17(0.2048)	42(0.0727)	59
	未検出数	66(0.7952)	536(0.9273)	602
	列-合計	83	578	661

$$\text{適合率: } \frac{50}{371} = 0.1348$$

$$\text{再現率: } \frac{50}{71} = 0.7042$$

$$\text{正解率: } \frac{50+382}{774} = 0.5581$$

正解率が約 56%と半分程度の値になった一方で，適合率が約 13%と低い値になった．正解率に対して，適合率が低いため，ツール出力には誤検出が多いことがわかる．

次に，RQ2 に回答するために，表 4 から，真陽性率と偽陰性率，偽陽性率と真陰性率を算出した．その結果を表 5 に示す．

表 5: 2 群の各割合

#	真陽性率	偽陰性率	偽陽性率	真陰性率
実験群	**0.7042**	**0.2958**	0.4566	0.5434
統制群	0.2048	0.7952	**0.0727**	**0.9273**

真陽性率が実験群は約 70%であったのに対して，統制群は約 20%であった．また，偽陰性率では，実験群は約 30%であったのに対して，統制群は約 80%であった．

この結果から，実験群は誤った要素の検出が統制群よりも網羅的に行えていることが確認できる．

6 考察

RQ1 と RQ2 から，ツール出力は誤検出を多く含むが，より多くの誤った要素を検出することが出来ることが確認できた．実験群のツール出力をカテゴリごとに分けた表 6 から，特に，要素の不足に関して，誤検出が多く，検出できなかった誤った要素も多いことが分かる．

表 6: カテゴリごとのツール出力の一致数

		妥当性評価結果			
	#	誤った要素	正しい要素	見逃し数	行-合計
実験群	要素の不足	22	185	17	224
	無関係な要素	3	51	2	56
	関連付けの誤り	5	22	0	27
	図形の誤り	16	59	1	76
	線の誤り	0	0	0	0
	要素の重複	2	4	0	6
	未入力	2	0	1	3
	列-合計	50	321	21	392

この理由として，要求文を最大 3 つの分節に区切ったテキストと，モデルの要素ごとのテキストで類似度計算を行って検出している．例として表 7 の場合，モデル要素の記述に対して，最も類似度が高い要求文の記述が 3 つの分節であるため，正しい類似度計算が行えていない．解決策として，モデル要素ごとの分節長に合わせて，要求文の分節を区切ることで誤検出を減少させることができる可能性がある．

表 7: 要素の不足の誤検出例

#	テキスト (分節長)
モデル要素の記述	受託者以外の者であっても同様のサービスを一般的な手段で調達することが可能 (9)
要求文の記述	受託者以外の者であっても (3)
類似度	0.6484

7 まとめ

本研究では，モデリング時に発生する誤りに対して，要求文の意味を考慮したうえで，機械的処理による自動発見を行うために，7 つに分類した誤りのカテゴリを作成し，誤り箇所の自動検出を行った．

実験結果から，提案手法によって，確実なモデルの誤り箇所を網羅的に検出できる一方で，正しい要素を検出してしまう誤検出が多いことが確認できた．

今後の課題として，誤検出の削減による正解率と適合率の向上，被験者に対する iStar の説明の改善もしくは学習時間の追加が挙げられる．

参考文献

[1] Fabiano Dalpiaz, Xavier Franch, and Jennifer Horkoff, "istar 2.0 language guide," May 2016.

[2] 総務省，"総務省デジタル・ガバメント中長期計画," `https://www.soumu.go.jp/main_content/000841514.pdf`, Oct. 2022.

[3] Yoshitake Hirabayashi, Shinji Ohota, Suzuka Fujii, and Motoshi Saeki, "Defining bad smells and automating their detection in goal-oriented requirement analysis method istar," Proc. of the 30th Asia-Pacific Software Engineering Conference (APSEC) 2023, pp.349–358, 2023.

[4] Charles H. Kepner and Benjamin B. Tregoe, The New Rational Manager, Princeton Research Press, Princeton, New Jersey, 1981.

[5] Jacob Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement, Vol.20, pp.37–46, Apr. 1960.

[6] Kilem Li Gwet, "Computing inter-rater reliability and its variance in the presence of high agreement," British Journal of Mathematical and Statistical Psychology, Vol.61, No.1, pp.29–48, May 2008.