レビュー解析を題材とした「誤り分析マニュアル」の試作に向けた検討

藤井 敦
東京工業大学

乾 孝司
筑波大学

中山卓哉
筑波大学

1. はじめに

自然言語処理の研究における評価実験の重要性が益々増加しており，研究課題によってはデータセットから実験結果の提示方法に至るまで，事実上の標準が存在する[1,2]．しかし，実験結果に対する考察とりわけ提案手法が有効に機能しなかった誤り事例に対する分析すなわち「誤り分析」は著者によって提示の有無や内容に差異がある．

誤り分析の動機，目的，方法が一通りであるとは限らない．ただし一つの見解として，手法の得失を理解し状況に応じて使い分けることや，誤りの原因を究明して手法を改善することが動機である．そのために，誤りの原因を類型化し，解決手段に関する検討を容易にすることが目的である．問題は，誤り分析の方法論が研究者間で共有されておらず，分析者による品質のばらつきが大きい点にある．

本稿は，誤り分析の品質向上を目的として，分析作業のマニュアル化に関する方法論を提案し，現状について報告する．レビュー解析における評価分類を題材として分析作業の体系化と定型化について研究を行いつつ，他の研究課題への応用も検討する．学生の研究指導や若手研究者の育成といった教育目的の利用にも意義があると考えている．

当マニュアルの趣旨は，誤り分析の下限を一定以上の水準に保つ点にあり，上限を設けることは意図していない．むしろ，独創的な分析事例の良い面を吸収してマニュアルを随時に発展させていきたい．

2. 誤り分析とは何か？何でないか？

誤り分析には未知の領域が多く，それが本研究の動機でもある．しかし，以降の議論を円滑に進めるために，現段階での誤り分析に関する著者らの見解を明記する．

評価実験は，条件ごとに正解率等の評価値を集計する作業を伴う．しかし，こうした機械的な集計では決して見えない真実を炙り出すことこそが誤り分析の本質である．言い換えれば，様々な評価値が整然と並んだ見栄えの良い「表」を量産する作業は誤り分析ではないと考えると分かりやすい．

なお，誤り分析の前に正反対の作業，すなわち成功事例が提案手法に起因するものか否かを分析して偶然の産物ではないことを確認しておくと良い．

誤り分析とは何か何でないかを以下にまとめる．

- **サーベイではなくリサーチである**．調査項目は事前に決まっていない．
- **黙思ではなく目視である**．夢想するより現場百篇である．
- **観察ではなく洞察である**．事実の背後に潜む真実を見抜く．
- **理想ではなく現実である**．都合の良い解釈をしない．
- **報告ではなく説得である**．証拠固めをする．例えば，誤り原因の候補を取り除くことで本当に問題が解決したか確認する．

3. 研究の方法

本研究は，誤り分析において分析者によって差異が生じる要因を特定し，それを解消することで分析作業をマニュアル化する．今回は，同じ実験結果に対して著者らが個別に誤り分析を行い，作業手順や分析結果の共通点と相違点について考察した．

実験結果とは，楽天トラベルのレビューから抽出した文に対して肯定，否定，中立の三値分類を実行した結果である．使用したデータは，筑波大学文単位評価極性タグ付きコーパス(TSUKUBA コーパス，http://www.mibel.cs.tsukuba.ac.jp/~inui/SA/corpus/)の一部であり，オリジナルデータは楽天データ公開（http://rit.rakuten.co.jp/opendataj.html）によって公開されている．

文数は 2700(肯定 1379，否定 639，中立 682)であり，サポートベクターマシンを用いて 10 分割交差検定を行った．以下にその他の実験条件を示す．

・単語ユニグラム素性
・線形カーネル
・ソフトマージンコストの値:1
・ペアワイズ法で多値分類

上記の手法(ung)とは別に，評価表現を素性に用いた手法(dic)も準備した．ung と dic の分類正解率

は，80.78%（2181/2700）と 81.59%（2203/2700）であった．今回は単純な手法を分析の対象とすることで，誤り分析のうち技術的な複雑さに起因する問題を分離した．また，基礎的な手法を対象とすることで大きな波及効果を狙っている．

誤り分析のために構成した本稿著者らのグループ二つをそれぞれ G1 および G2 と呼ぶことにする．分析の方針等に関する事前の調整はせずに独立に誤り分析を行い，結果を比較した．

結論から言えば，どちらのグループも誤り事例を誤りの原因に基づいて類型化した点は共通であった．しかし，類型化の過程において両グループはカテゴリ分類とクラスタリングという異なる方法を用いた．それにもかかわらず，両グループの分析結果には，いくつかの共通点があった．以下の 4 章と 5 章で，各グループの分析方法と結果について説明する．

4. G1 グループの分析方法と結果

評価分類の研究実績に基づいて，評価表現に関する誤り原因の類型を事前に予想し，誤り事例を類型に振り分けた．使用した類型を以下に示す．

A： 正解極性の潜在的な評価表現が存在する（例：イマイチ）．潜在的評価表現が把握できれば改善が期待できる．潜在的評価表現は，大抵の場合は直接的な表現である．

B： 正解極性の潜在的な評価表現が単語の組合せとして存在する（例：おなかいっぱい）．当該評価表現が把握できれば改善が期待できる．大抵ある特定の評価を喚起させると推測できる事態の記述である．

C：文中に肯定と否定が混在している．

D： 正解極性の既知評価表現があるにもかかわらず誤った．

E：上記以外の原因による．

A～E の複数が該当する場合は以下の優先順位に従って一つを付与した．

C > A = D > B > E

G1 内の二名が独立に作業を行い，結果を比較した表を次に示す．

	A	B	C	D	E
作業者 1	115	97	16	35	234
作業者 2	112	94	16	31	244

表では事例の分類クラスが示されていないが，本類型は評価表現を軸にしていることから，中立が正解である誤り事例の多くは E に分類され，実際，それらが E の過半数を占めていた．

作業者間で A～E の振り分け先が一致しなかった主な理由は以下の通りである．

・A, B: 単語か複合語を決める基準の違い．
・A, E: 評価表現として認める基準の違い．
・B, E: 単語の組合せとして認める基準の違い．

なお，G1 は ung と dic の比較に基づいて dic の誤り分析を試みたものの，両手法の結果に大差がなかったため，この点に関する詳細は割愛する．

5. G2 グループの分析方法と結果

事前の予想や仮説を持たずに誤り事例を一つずつ見ながら誤りの原因を徐々に一般化し，この作業を反復しながら類型を統廃合していった．ただし，何らかの共通点を持った事例を集中的に見るために，正解と自動分類のパターンごとに分析を進めた．

例えば，肯定と否定が相互に誤分類される場合は，正解に特徴的な単語の不足もしくは不正解に特徴的な単語の過剰が原因であることが多いのに対して，中立が関与する場合はそもそも中立に特徴的な単語が少ないため他の原因が主流であった．

結果的に，表 1 に示すような三階層の分類体系が作成された．評価表現に関する誤り事例は G1 の分析結果と共通点が多かった．他方において，手法の狙いであった単語素性に起因する誤りが細分化されている点や，文間関係のように手法が狙っていない解決手段を示唆している点が異なる．

6. マニュアル化に向けた展望

作成過程や粒度に違いはあるものの，G1 と G2 の類型化は「本来の狙いが外れた原因」を特定することを意図している．すなわち，評価対象の手法が明示的もしくは暗黙的に依存している仮定を見極めることが当該手法の潜在的な弱点の発見につながる．

以下，今回の誤り分析実験から得られた知見を踏まえて，本研究が目指すマニュアルの設計方針に

ついて議論する．まず，マニュアルのモデルとして，チュートリアル，リファレンス，トラブルシューティング，用語集を想定する．

チュートリアルは，教科書のように通読や演習を通して体系的な基礎知識を与える素材である．誤り分析実験で用いたデータや得られた知見を利用することができる．

リファレンスは，誤り分析の最中に見つけた特定の事例をきっかけとして調べるための素材である．チュートリアルが最初から通読することを前提としているのに対して，リファレンスは索引のように特定の語句による逆引きを可能にする．誤りの原因が付与されたレビュー文の事例集合から，自分が分析中の事例と類似する事例を検索する機能が有効である．

トラブルシューティングは，本稿著者らが経験した誤り分析の「落とし穴」と脱出方法に関する事例を蓄積し，提供することを検討している．

用語集は，チュートリアル，リファレンス，トラブルシューティングに出現する用語の解説である．レビュー解析に関する専門用語や誤り分析に関する概念などについて，ウェブ等を検索できるように一覧しておく．

7. おわりに

レビュー解析を題材とした誤り分析のマニュアル作りを目的として，分析における個人差について考察するとともにマニュアル作成の指針について議論した．今後は，さらなる検討を重ねてマニュアルを試作する予定である．

8. 謝辞

データを提供頂いた楽天トラベル株式会社および関係諸氏に深く感謝いたします．

参考文献

[1] E. M. Keen. Presenting results of experimental retrieval comparisons. Information Processing & Management, 28(4):491–502, 1992.

[2] M. Sanderson. Test collection based evaluation of information retrieval systems, Foundations and Trends in Information Retrieval, 4(4):247–375, 2010.

表1：評価分類タスクに関する誤り事例の分類体系

大分類	中分類	小分類	具体例
狙いが外れた	正解の特徴語が力不足	評価表現	肯定:「おいしい」 否定:「今ひとつ」
		表記ゆれ	肯定:「有難い」や「有り難い」は代表表記でない
		未知語	否定:「バサバサ」
		誤記	肯定:「気に入る」を「気に入れる」と誤記
		定型句	肯定:「気を利かす」 否定:「～してほしい」
		特殊記号	肯定:「◎」 否定:「．．．」
		修辞疑問	否定:「～があっても良いのではないでしょうか？」
		学習データ	疎問題やデータ偏向
		特徴語なし	中立に多い
	不正解の特徴語が過剰	頻出語	肯定:「とても」 否定:「ただ」
狙っていない	参照表現		「バス、トイレなしの予約でしたが、両方ついたお部屋」の「両方」が「バス、トイレ」を指す
	文間関係		全体的に肯定か否定に傾倒　極性の継続（「特筆すべきは」）や反転（「しかし」）
	領域知識		肯定:「3回目の宿泊」はリピーターを示唆
	比較		否定:「料金の割にせまい」
	仮定		否定:「露天風呂があれば良かった」