E-030

小説検索のための同義タグの認識に関する一考察
A Study on Synonymous Tag Recognition for Novel Retrieval

鈴木 信太郎 * 坂野 妃菜 * 谷口 雅空 * 工藤 竜矢 *
Shintaro Suzuki Hina Sakano Masataka Taniguchi Ryuya Kudo
宍戸 伶雅 * 酒井 柊輔 * 延澤 志保 *
Ryoka Shishido Shusuke Sakai Shiho Hoshi Nobesawa

1. 研究背景

小説投稿サイトのひとつである小説家になろう[1]に投稿されている作品数は 2022 年 5 月 24 日現在 955,936 件と膨大である．読者が作品を検索する際に使用できる要素は，作品タイトル，あらすじ，タグ，作者名，ジャンルであり，一般的に検索に用いられる要素はタグである．タグは運営が主に著作権の管理のために設定したものと，作者が自由記述で登録したものの 2 種類に大別される．自由記述で登録されたタグは作品の絞り込みに用いる要素だが，多くの作品に共通して用いられているタグは少なく，目的にあった作品を検索することに適したタグを選択することは難しい [1]．あらすじの部分一致検索を行うことも可能だが，あらすじに記載される内容はさまざまであり，主な検索の要素としては不十分である．

2. 小説タグの特徴

本研究では，なろう小説 API[2]を用いて取得した小説情報を対象として，検索クエリとしてのタグについて検討する．2021 年 1 月 1 日までに投稿された小説を対象に情報を取得し，合計 711,749 作品分のデータを得た．この 711,749 作品に付与されたタグの異なり数 332,006 個のうち出現数 2 以上のタグは 76,839 個で全出現タグの 23.1% であり，全出現タグの 76.9%に当たる 255,167 種類のタグは 1 つの作品のみに付与されたものである．出現作品数が 10 以上のタグは 14,722 個でタグ全体の 4.4%に過ぎない．これは個々のタグが適切に小説を絞り込めることを示すわけではなく，タグの決定が作者に委ねられているためタグ文字列の統一が取れていないことに起因する．図 1 に全出現タグの文字数ごとの分布を示す．図 1 の横軸はタグの文字列の長さ，左縦軸はそれぞれの文字数のタグの異なり数，右縦軸は 10 作品以上に使用されたタグの割合を示す．小説家になろうで使用可能なタグの文字数上限は現在 10 文字で，11 文字以上のタグは 2009 年以前もしくは 2015 年の限られた期間に投稿された小説にのみ使用されたものであるため，11 文字以上のタグの異なり数は極端に少ない値となっている (図 1)．

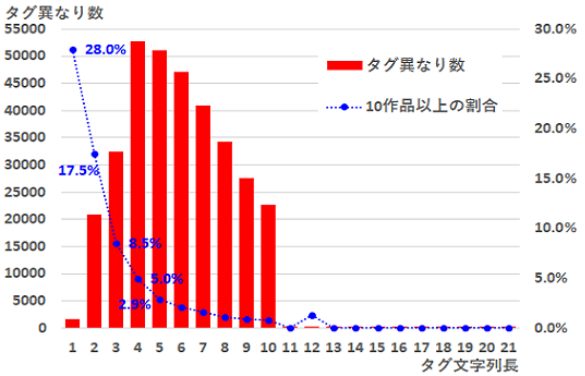

図 1: タグの文字列長とタグ数の関係

10 作品以上に使用されているタグの比率は 1 文字のタグでは 28.0%と比較的高いが，3 文字のタグになると 8.5%まで減少し，4 文字では 95.0%のタグが出現作品数が 10 に満たないタグとなる (図 1)．日本語の名詞の多くは漢字 2〜3 文字であり，創作小説のため固有名詞などでカタカナの割合が高いことを考慮しても，文字数 5 を超えるタグが半数を超える (図 1) のは単語ではなく複合語や句の形のタグが多いためと考えられる．複合語句からなるタグは作品の絞り込みには向かず，タグによる小説検索のヒット数の向上を目指すには類似のタグを互いに関連付ける必要がある．

3. 同義タグの種類

タグの類似性の推定の基盤として，本稿では，小説投稿サイトへ投稿されたオンライン小説に付与されたタグについて，同義とみなすことのできるタグの推定について検討する．

表 1 に意味が近い可能性があるタグの組の例を挙げる．表 1 の数値はそれぞれのタグが付与された作品数，

表 1: 意味が近いと考えられるタグの組合せの例

タグ 1 タグ	(作品数)	タグ 2 タグ	(作品数)
コメディ	(14,617)	コメディー	(12)
日常	(132,154)	ありふれた日常	(32)
先生	(1,026)	先生と生徒	(105)
悪魔	(3,979)	デビル	(10)
戦闘	(3,523)	バトル	(10,398)

下線は組合せたタグの共通文字列を表す．表 1 から，類義タグの組合せの種類として，表 2 を仮定する．表

表 2: 類義タグの組合せの種類

同義 a.	表記揺れ	同じタグの表記の違い
同義 b.	修飾	タグに対して修飾句を付加
同義 c.	異表記	同義語句の異表記
包含	上位下位概念	
類似	類義語句	

2 の同義タグ a は「コメディ」と「コメディー」などの表記揺れに起因する同義タグを示す．同義タグ b は「日常」に対して「ありふれた日常」のように修飾句な

* 東京都市大学，Tokyo City University
[1]小説家になろう，https://syosetu.com/.
[2]なろう小説 API，https://api.syosetu.com/novelapi/api/.

どを付加したもののうち，同義とみなせるものを指す．同義タグ c は「悪魔」と「デビル」など，異表記だが同義とみなせるタグを示す．これらは，例えば「ゲーム」と「乙女ゲーム」のような包含タグや，「男主人公」と「女主人公」のような類似タグとは異なり，同じ事物を表現することを目的としたタグと考える．

4. 同義タグの推定

本研究では，タグ同士の同義性の判定を目標として，タグ文字列とタグに対応する小説本文からタグの特徴を推定する手法を検討する．

本稿では，同義タグの推定の基礎となる尺度を検証するため，(1) タグに対応付けられた本文の類似度，(2) タグ文字列の類似度の 2 種類の尺度でタグの比較を行った．

(1) 本文に基づく類似度は，対象のタグそれぞれについて対応付けられた小説すべての本文から作成したベクトル同士の比較によって推定する．本稿では対応する小説に含まれる名詞および動詞の文書頻度 [2] をベクトルの素性とする．また本稿では，ベクトルの比較にコサイン類似度 [2] を用いる．

(2) タグ文字列に基づく類似度はレーベンシュタイン距離を用いて比較する．本稿では，タグ文字列間のレーベンシュタイン距離を 2 個のタグのうち長い方のタグの文字数で割ることで正規化する．さらに，本文類似度との比較のため，求められた正規化済みレーベンシュタイン距離を 1 から減じることで，レーベンシュタイン距離の小さいものほど 1 に近い値となるように調整する．

小説家になろうに 2017 年に投稿された 10,000 文字以上の小説 1,000 作品について，これらの作品の中で 10 作品以上に付与されていたタグ 114 個の組合せごとのタグ類似度を，本文類似度，文字列類似度の 2 つの尺度で比較した．図 2 に，文書の類似度としてコサイン類似度，タグ文字列の類似度として編集距離類似度の相関を表す．図 2 の横軸は文字列類似度，縦軸は本

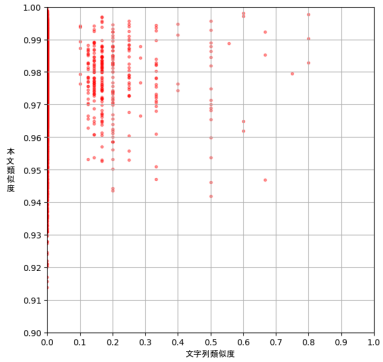

図 2: タグ文字列の類似度と本文の類似度の相関

文類似度を示す．図 2 に示すとおり，本文類似度はすべての組合せで 0.9 を超えており，本稿で提案した名詞と動詞の文書頻度に基づく類似度では十分な比較ができないことがわかる．

表 3(左) に本文類似度の高いタグの組合せのうちコサイン類似度が 0.998 を超えたもの 21 件，表 3(右) に文字列類似度の高いタグの組合せ上位 21 件示す．タグの赤い文字は組合せたタグの共通文字列を示す．

表 3: タグ類似度の高いタグの組合わせ

タグ1	タグ2	本文	文字列	タグ1	タグ2	本文	文字列
R15	残酷な描写あり	1.000	0.000	異世界転移	異世界転生	0.998	0.800
魔法	冒険	0.999	0.000	ネット小説大賞六	ネット小説大賞六感想	0.990	0.800
冒険	R15	0.999	0.000	コメディ	コメディー	0.983	0.800
魔法	異世界転移	0.999	0.000	男主人公	女主人公	0.979	0.750
魔法	R15	0.999	0.000	未来	近未来	0.992	0.667
冒険	残酷な描写あり	0.999	0.000	ボーイズラブ	ガールズラブ	0.985	0.667
冒険	異世界転移	0.999	0.000	少女	美少女	0.947	0.667
冒険	異世界	0.999	0.000	異世界	異世界転移	0.998	0.600
残酷な描写あり	オリジナル戦記	0.999	0.000	異世界	異世界転生	0.997	0.600
魔法	異世界	0.999	0.000	ミステリー	ミリタリー	0.965	0.600
魔法	残酷な描写あり	0.999	0.000	ゲーム	乙女ゲーム	0.962	0.600
R15	オリジナル戦記	0.998	0.000	星球大賞２	星球大賞２感想希望	0.989	0.556
R15	異世界転移	0.998	0.000	魔法	魔王	0.996	0.500
日常	ほのぼの	0.998	0.000	バトル	異能力バトル	0.993	0.500
青春	現代	0.998	0.000	魔法	剣と魔法	0.989	0.500
冒険	男主人公	0.998	0.000	ゲーム	ハーレム	0.988	0.500
異世界転移	残酷な描写あり	0.998	0.000	異世界	電脳世界	0.986	0.500
冒険	オリジナル戦記	0.998	0.000	ガールズラブ	スクールラブ	0.984	0.500
異世界	R15	0.998	0.000	恋愛	古典恋愛	0.982	0.500
異世界	異世界転移	0.998	0.600	戦闘	戦争	0.975	0.500
R15	異能力バトル	0.998	0.000	私小説	探偵小説	0.971	0.500

表 3(左) を見ると，「Ｒ１５」と「残酷な描写あり」の組合せのように，文字列は異なるが表現する事物は同様と考えられるものがあり，文字列類似度だけでなく本文類似度による比較が有効と考えられる．そのため，名詞および動詞の文書頻度に基づく本文類似度よりも信頼性の高い本文類似度の算出手法を検討することで，同義性推定の精度向上が期待できる．

表 3(右) では「異世界転移」と「異世界転生」のように，タグ文字列に共通部分の多いタグ同士の組合せは内容も類似している可能性が高いことがわかる．例えば「異世界転生」は，本文中にこの語が頻繁に出現することは考えにくく，本文類似度だけでなく文字列類似度が同義性の推定に寄与する可能性を示す．半面，文字列類似度の高い組合せ (表 3(右)) では「未来」と「近未来」のような包含タグ，「男主人公」と「女主人公」のような類似タグも含まれており，文字列類似度だけでは同義性を正しく推定できない場合があることがわかる．特に，「ミステリー」と「ミリタリー」のように偶然共通文字列を多く含む場合があることを考えると，文字列類似度の算出方法についても，検討の余地がある．

5. まとめ

本稿では，小説投稿サイトの主な検索の要素であるタグの同義性の推定について検討した．本稿では，タグの同義性の推定尺度として本文類似度と文字列類似度の 2 種類を提案した．小説 1,000 作品に対して 10 作品以上に付与されていたタグ 114 個についてそれぞれの組合せの同義性を推定した結果，本文類似度，文字列類似度はそれぞれに同義性推定の観点が異なり，双方を組合せることでタグの同義性の推定が期待できることが示された．

参考文献

[1] 蒲生 奏衣, “類似度に基づく Web 小説のタグ推薦,” 東京都市大学学位論文, 2022.

[2] Christopher D.Manning, Hinrich Schutze, “統計的自然言語処理の基礎,” 共立出版, 2017.