書誌情報と主題語彙のリンキングによるマンガの主題検索のためのデータセット構築

A Dataset Development for Subject Searching from Metadata and Vocabulary about Subject of Manga

志賀大輝[1]　三原鉄也[2]　永森光晴[2]　杉本重雄[2]

Daiki Shiga[1], Tetsuya Mihara[2], Mitsuharu Nagamori[2] and Shigeo Sugimoto[2]

[1]筑波大学情報学群情報メディア創成学類

[1]College of Media Arts, Science and Technology, School of Informatics, University of Tsukuba

[2]筑波大学図書館情報メディア系

[2] Faculty of Library, Information and Media Science, University of Tsukuba

Abstract: Descriptive metadata, mainly subject is used to provide search within the vast collections of Manga on the web. Manga subject vocabularies are created to improve the metadata used for subject-based searches. However, considering the large number and diversity of subjects covered, developing such a subject vocabulary for Manga is a tedious process. This research proposes a method to enhance subject vocabulary by expanding the subject headings to improve the inclusiveness. The proposed method initially links existing vocabulary with the metadata obtained from major e-book providers, and then extend subject headings with theme-related words extracted from other web resources.

1.はじめに

現在、日本において多数のマンガが出版されており、2018 年版出版指標年報によると、2017 年に出版されたマンガの新刊点数は 12,461 点にも上る[1]。この膨大なマンガの中からユーザが自身の要求に合致するマンガにアクセスできるようにするために、電子書籍ストアでは、作品タイトルや著者、出版社などのメタデータを利用する検索サービスを提供している。特にマンガの主題に関するメタデータが主題語である。これを用いることでユーザはマンガを読まなくても、自身の嗜好に合致した主題を持つマンガにアクセスすることができる。

しかし、現状ではメタデータとして与えられる主題語はシンプルなテキストによるキーワードであることが多い。そのため主題語を利用した検索はその文字列の比較によるものに留まり、図書館の分類や件名標目のような主題語の表す意味や主題語同士の関係に基づいた検索を行うことができない。主題語の表す意味に基づいた検索を行うためには、主題語の語彙を構築する必要がある。

主題検索を行うためには、マンガに付与された主題語によって、その主題の差異を区別できる必要がある。しかし、新刊の出版などによってマンガの数は変化するため、必要な主題語の数は一定ではない。そのため、マンガの主題検索に適した主題語の語彙を構築するためには継続的にその拡充と整備をする必要がある。しかし、一般に主題語の語彙の構築は人手によって行われるため、膨大に存在するマンガの主題の差異を表現するための大規模な語彙の構築は難しい。

そこで本研究では、人手で構築したマンガの主題語彙においてマンガの主題の差異を区別できていない箇所の発見と、そこに拡充する主題語の候補の提示を行うことで、マンガの主題語彙の構築を支援する。そうすることで、マンガの主題検索に適した主題語彙の構築を行う。

2.マンガの主題検索のための語彙構築

ユーザが自身の嗜好に合う書籍にアクセスする際には、タイトルや著者名などをはじめとした、書誌情報を利用することが考えられるが、特にその内容に即して探索したい場合には、ユーザに対して書籍の内容に関する情報が提供されている必要がある。その代表的な方法として、日本十進分類法があげられる[2]。これは、「哲学」「歴史」「自然科学」などの内容に基づいて書籍を分類することで、ユーザが求

める内容を持つ書籍へのアクセスを可能としている。電子書籍ストアでも同様に、取り扱う書籍に合わせた分類を用意することで、書籍の内容に関する情報を提供している。

しかし、日本十進分類法のようにマンガを内容によって分類するのは難しい。なぜならば、マンガのテーマや作風による分類は、多かれ少なかれ個々人の主観に左右され、分類体系として広く共有されるとは言いがたいためである[3]。そのため電子書籍ストアでは、マンガのテーマや舞台、登場人物など、マンガをさまざまな観点から見たときの主題語を付与している。これは例えば「恋愛」や「ラブコメ」、「学校」といった語である。これを利用することで、ユーザは主題語をクエリとして入力し、自分の求める主題をもつマンガを検索できる。

現状では、マンガに付与される主題語はシンプルなテキストによるキーワードであることが多い。よって、主題検索は文字列の比較によるものに留まり、図書館の分類や件名標目のような主題語の表す意味や主題語同士の関係に基づいた検索を行うことができない。この問題は、主題語の表す意味的な関係を記述した語彙を構築することで解決することができる。

マンガの主題語は、その語のもつ一般的な意味とは異なり、マンガの文脈特有の意味をもって主題を表現していることや、物語中で登場する架空の職業名や生物名などが主題語として扱われることがある。そのため、マンガの主題語彙の構築を行う場合に、件名標目などの既存の語彙を利用して主題語の関係を記述することが難しく、その構築は人手によって行う必要がある。また、主題語を付与する対象となるマンガの数が新刊の出版などの要因によって変化するため、その語彙の拡充と整備を継続的かつ頻繁に行う必要がある。

しかし、マンガの主題検索を行うには、膨大な数のマンガの主題の類似や差異を主題語によって区別できる必要があるため、その語彙は大規模になる。そのため、マンガの主題語彙の構築する場合に、その拡充と整備を人手によって行うことは負担が大きい。

そこで本研究では、マンガの主題検索に適した主題語彙を構築することを目的とし、語彙の拡充と整備を支援する手法を提案する。前述したように、主題語を利用してマンガを検索するには、それによって主題の類似や差異を表現できる必要がある。したがって、多数のマンガに付与されている主題語、あるいは逆にほとんどマンガに付与されていないような主題語ではマンガを検索する際に利用し辛いといえる。これらの課題を改善することで、マンガの主題検索に適した主題語彙を構築することが可能になると考えられる。そこで本研究では、主題語彙中の主題語がマンガにどのように付与されているかに着目し、その改善の必要な箇所への主題語の拡充を支援する。

3.関連研究

コンテンツの主題に基づくアクセスを行うためには、コンテンツから主題情報を抽出する必要がある。馬場らはWeb上の小説を対象に主題情報の抽出を行った[4]。この研究では小説の「ジャンル」と「登場人物」による分類を目的としている。形態素解析や既存の辞書を利用して特徴量となる語を抽出し、小説のジャンル推定と人物抽出の手法を提案している。しかし、小説の主題に関する情報はテキストから得られるのに対して、マンガはテキストとイラストが混交してストーリーを構成するマルチモーダルなコンテンツであり、画像処理技術等によってマンガそのものからその主題情報を機械的に抽出することが難しい。また、膨大な数存在するマンガの主題の情報を人手で抽出するのは現実的ではない。

そこで、マンガについて記述された外部の情報源から主題を表現する情報を抽出し、マンガの主題に基づくアクセスを目指す研究が行われている。山下らは、レビューサイトにおいてマンガについて記述されたレビュー文を分析対象とし、レビュー文中から抽出した単語について hLDA(hierarchical Latent Dirichlet Allocation)のトピックモデルによってトピック分類を行うことによりマンガ作品に含まれるトピックを推定し、トピックを共有するマンガの提示などによりマンガの内容情報に基づくアクセス支援を行っている[5][6]。しかし、トピックを共有するマンガの検索に留まり、マンガのトピック同士の意味的な関係の記述は行われないため、類似する主題の検索など、主題の意味的な関係に基づいた検索を行えない。

本研究は、主題語彙の構築を行い、主題語の表す意味に基づいたマンガの検索を目指す点で山下らの研究とは異なる。

4.電子書籍ストアのメタデータを用いた主題語彙の拡充

4.1 主題語彙の拡充手法

２章で述べたように、マンガの主題語彙を構築する場合、その拡充と整備を人手で行うのは難しい。

そこで本研究では、マンガの主題語彙における拡充の必要な箇所の発見と、拡充する主題語の候補の提示を行うことで主題検索に適したマンガの主題語彙の構築を支援する。

そのために、まず、主題語彙をマンガと結びつけることで、人手で構築したマンガの主題語彙において、多くのマンガに付与されている主題語など、マンガの主題の差異を表現できていない箇所を発見する。次に、Web 上の情報源を利用して、主題語彙を拡充するための語の候補を取得する。この２つの手順によって、マンガの主題語彙の拡充を行う。

マンガの主題語彙とマンガの結びつけを行うためには、マンガの主題情報を取得する必要がある。本研究では、大量に存在するマンガの主題についての情報を得るために、電子書籍ストアで提供されている主題のメタデータを用いる。これは電子書籍ストアごとにタグ、ジャンルなど様々なプロパティで表現されるが、本稿では以降タグと呼称する。タグを用いることでマンガを読むことなくその主題情報を取得できるため、大量のマンガの主題に関する情報を取得できる。そこで本研究では、電子書籍ストアからマンガとそこに付与されるタグについての情報の抽出を行う。

4.2 書誌情報とのリンキングに基づく拡充箇所の発見

マンガの主題語彙において、マンガの主題の差異を表現することができていない箇所を発見するために、マンガに付与されたタグを利用して主題語彙とマンガの結びつけを行う。これはタグと主題語のリンキングと、リンキングの結果を利用して主題語彙にマンガを結びつける２つの作業によって行う。

まず、タグと主題語のリンキングを行う。本研究では、主題語とタグの間の関係を次の５つのうちのいずれかとした。どの関係にあたるかを決定する処理は人手によって行う。

(1) 主題語とタグが同じ概念を表す
(2) タグが主題語の別表記あるいは同義語・類義語である
(3) タグが主題語の広義の概念である
(4) タグが主題語の狭義の概念である
(5) タグと主題語の間に何らかの関連がある

次に、マンガに付与されたタグがもつ主題語へのリンクに基づいて、マンガを主題語彙に結びつける。例えば、タグ「お風呂・温泉」が付与されたマンガは、その主題語によって図 1 のようにして主題語彙中の語「お風呂」「温泉」「温泉旅館」にそれぞれ結びつけられる。その結果から、主題語彙中の主題に対しマンガがどのように結びついているかを見ることで、主題語彙において拡充の必要な主題の発見を行う。

ただし、各電子書籍ストアで販売されるマンガの単位はそれぞれ異なるため、そのマンガの単位で結びつけを実行すると、主題語彙に結びつくマンガの単位がそれぞれの電子書籍ストアで提供されているものになってしまう。その場合、結びつけられるマンガの単位が一定でなくってしまうため、電子書籍ストアで販売されているマンガの単位で主題語彙との結びつけを行うことは適切ではない。そこで、本研究では、電子書籍ストアで販売されている単位をマンガとして扱うのではなく、各電子書籍ストアにおいて、複数巻・複数話に分けて販売されるマンガを集約した単位をマンガとして扱い、主題語彙との結びつけを行う。

4.3 主題語の候補の取得

前述したように、マンガの主題語彙を構築する場合に、主題語の関係の記述は人手によって行う必要がある。そこで本研究では、主題語彙へ追加する語の候補を提示することで、その構築を支援することができると考え、それをマンガのレビュー文と既存の外部語彙の２つの Web 上の情報源から取得する。

(1)マンガのレビュー文の利用

電子書籍ストアで提供されるマンガのレビュー文からの主題語の候補の取得を行う。レビュー文は、その多くがマンガに関する感想や批評などであり、そこにはマンガの主題に関する情報が記述されていると考えられる。そのため、あるタグが付与された

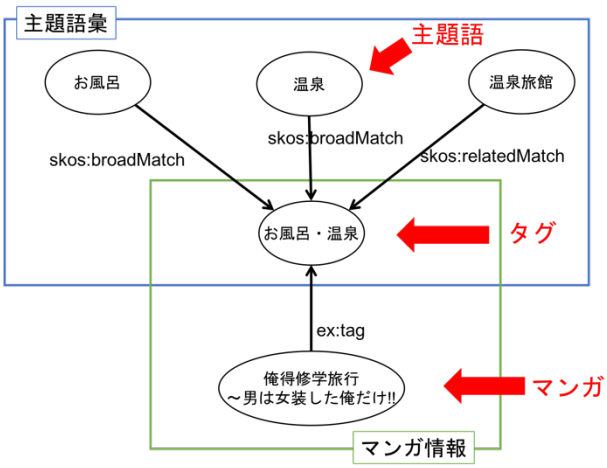

図 1 マンガと主題語彙の結びつけ

マンガ作品のレビュー文中には、そのタグと何らかの関係のある語が存在している可能性があると考え、レビュー文から主題語の候補の取得を試みる。

マンガのレビュー文からの主題語の候補の取得手法を述べる。図 2 は手法の模式図である。まず、マンガのレビュー文に対して形態素解析を行うことで単語を切り出す。それをレビューが記述された対象のマンガに付与されたタグごとにまとめ、あるタグが付与されたマンガ全てのレビューから切り出した単語の集合とする。その集合を、他の主題語の付与されたマンガのレビューに現れる単語と比較する。これにより、あるタグの付与されたマンガの全レビュー文中の単語の集合に見られる特徴的な単語、すなわち特徴語を取り出す。特徴語の抽出はレビュー文に対して TF-IDF を適用することによって行う。TF-IDF は文書中に含まれる単語の重要度を評価する手法の 1 つであり、その特徴量は文書中の単語の出現頻度 TF(Term Frequency)と逆文書頻度 IDF(Inverse Document Frequency)の積によって表される。

以上の方法を用いて取得した特徴語から、タグと主題語の間に作成したリンクに基づいて、ある主題語に結びつくマンガのレビュー文に現れる特徴語を取得する。この結果において特徴量の高い単語から主題語の候補の取得を行う。

(2)外部語彙の利用

既存の外部語彙において記述されている関連を利用して主題語の候補の取得を行う。本研究では、ユーザが記事を自由に作成・編集することができるオンライン百科事典を外部語彙として利用する。オンライン百科事典には様々な事象についての記事が存在し、それらは記事を作成したユーザの解釈に基づいて記述されるため、単語が表現する一般的でない概念やそれに基づく他の単語との関係、架空の単語の記事などが見られる。そのため、マンガの主題語に見られる、単語の一般的な意味とは異なる、マンガの文脈特有の意味が記述されていることが期待できる。

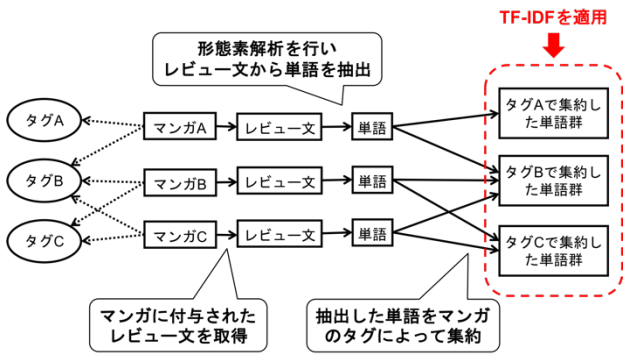

図 2 レビュー文からの主題語候補の取得手法

外部語彙からの主題語の候補の取得手法について述べる。はじめに、主題語彙の中の主題語をクエリとして外部語彙に問い合わせ、主題語の単語記事が存在するかを確認する。単語記事が存在するものについてはそれを参照し、そこに記述されている他の単語記事の情報を取得する。そして、外部語彙中で定義されている関係に基づいて単語記事を取捨選択し、主題語に関連すると考えられる単語を取得する。その結果から、マンガ向け主題語彙へ追加する候補の語を取得する。

5.データセット構築実験

5.1 実験に用いた主題語彙の概要

本研究では、マンガと結びつける主題語彙として、電子書籍ストア事業者が配信マンガのタグの構造化のために開発した語彙を提供頂き、これを利用した。この語彙は特に、その読者に成人を対象とするジャンル「オトナコミック」のマンガに主題語を付与するためのものである。これは、「ジャンル」「肩書き」「身体」「人間関係」「舞台」「性格」「服装」「状況」「合意」「プレイ」の 10 個の観点から主題語を構造化することでマンガの主題を表現しており、電子書籍ストアでマンガを提供する専門家の手によって構築されたものである。表 1 に示すのが、「舞台」の観点から構築した主題語彙の一部である。「category」の列に「ジャンル」「舞台」などの 10 個の観点のいずれかが入り、「subject」の列の語が主題語彙で扱う主題語を表し、「keywords」の列の語がマンガに付与する主題語を表す。また、右側の列の subject あるいは keywords の主題語は、1 つ左側の列の category あるいは subject の主題語に対して狭義の主題や関連を持った主題を表現する。例えば、「恋愛」「恋物語」「ラブストーリー」「オフィスラブ」「ラブコメディー」は主題「恋愛」をもつマンガに付与されることが考えられる主題語である。

この主題語彙に対して、10 個の観点ごとに同義語・類義語の関係にあたる主題語をまとめ、主題語の優先ラベルと別名ラベルを決定する処理を行い、854 件の主題語をもつ主題語彙を構築した。

5.2 電子書籍ストアからの情報抽出結果

本研究で用いるマンガの情報と結びつける主題語彙は、特にジャンル「オトナコミック」向けに構築

表 1 主題語彙の例

category	subject	keywords
ジャンル	恋愛	恋愛,恋物語,ラブストーリー,オフィスラブ,ラブコメディー
	学園	学園
	オフィス	オフィス,サラリーマン

表 2 電子書籍ストアからの情報抽出結果

	ストア A	ストア B	ストア C
マンガ作品	16,370	30,267	9,788
マンガ	184,689	237,282	27,528
レビュー	189,788	53	29,435
タグ	108	153	83

されたものである。これは、対象とする読者に基づく分類である。これを踏まえ、情報の抽出は、電子書籍ストアにおいて同様の読者を対象としていると考えられるジャンルに分類されているマンガに対して行う必要がある。したがって本研究では、マンガにタグが付与されており、かつ「オトナコミック」と同様の読者を対象とするマンガを扱っている電子書籍ストアから、取り扱うマンガの数や事業規模等を考慮し選択した３つの電子書籍ストア A[*1]、B[*2]、C[*3]から情報を抽出した。情報の抽出は、各電子書籍ストアにおいて「オトナコミック」と同様の読者を対象とするマンガとそこに付与されるタグ、およびレビュー文を対象に、Web スクレイピングによって行った。

情報を抽出した結果が表２である。抽出した情報の項目について説明する。マンガ作品とは、各電子書籍ストアにおいて、複数巻・複数話に分けて販売されるマンガを集約した実体のことである。マンガとは、1 巻・１話といった電子書籍ストアにおけるマンガ作品の販売単位を１つの実体としたものである。レビューとは、それぞれの電子書籍ストアにおけるマンガ作品に対して記述しているレビューである。タグとは、それぞれの電子書籍ストアで使用されているタグである。

5.3 書誌情報と主題語彙のリンキング結果

はじめに、マンガの書誌情報であるタグと主題語彙のリンキングを行った。タグには、マンガの主題を表現するものの他に、マンガが受賞した賞に関する語、マンガの形態についての語、マンガのメディア展開に関する語、特定のシリーズを表す語などが使用されている。これらのタグは、マンガの主題を表現する語とはいえないため、これらに対してリンキングを行うのは適切ではない。よって、はじめにこういった語を結びつけに使用するタグのリストから排除した。排除後のタグの数は電子書籍ストア A で 99 件、B で 120 件、C で 73 件の計 292 件となった。

表 3 タグと主題語のリンキング結果

関係	リンクの数
(1)	159
(2)	42
(3)	29
(4)	126
(5)	95
計	451

以上の処理を行った後のタグについて、主題語彙の主題語とのリンキングを行った。リンキングの結果は、タグと主題語の間にリンクを作成することで表現する。そのために、件名標目やシソーラスなどの統制語彙の構造を表現するためのモデルを提供する SKOS(Simple Knowledge Organization System)で定義されるプロパティを使用した[7]。使用したプロパティは、異なる概念体系のマッピングを記述するために用いられるもので、幅広い情報検索アプリケーションにまたがって交換して使用できるほどに同一であるといえる概念を結ぶ skos:exactMatch、一部の情報検索アプリケーションで交換して使用できるほどに同じであるといえる概念を結ぶ skos:closeMatch、２つの概念間の階層マッピング・リンクを記述する skos:broadMatch と skos:narrowMatch、概念間の関連マッピング・リンクを記述する skos:relatedMatch の５つである。

4.2 で示した、本研究で扱う関係を表現するプロパティは、(1)の関係を skos:exactMatch、(2)の関係を skos:closeMatch、(3)の関係を skos:broadMatch、(4)の関係を skos:narrowMatch、(5)の関係を skos:relatedMatch とした。その結果を関係ごとに表したものが表 3 である。

次に、リンキング結果に基づいて、マンガと主題語彙とを結びつけを行った。結びつけは、プロパティ skos:exactMatch を用いて行った。

結びつけの結果、結びついたマンガの数が多かった主題語の上位 5 件は表４のようになった。最も多くのマンガが結びついた主題語では、4,784 件のマン

[1] https://sp.comics.mecha.cc/

[2] https://sp.handycomic.jp/

[3] https://www.cmoa.jp/

表 4 主題語彙とマンガの結びつけ結果

主題語	マンガの数
人妻	4,784
調教	3,850
複数プレイ	3,705
幼なじみ	2,215
巨乳	1,959

ガに付与されており、15 件の主題語が 1,000 件を超えるマンガに結びついていることがわかった。また、マンガに全く結びつけられなかった主題語も多数存在する。今回、構築された主題語彙の主題語 854 件のうち、 1 件以上のマンガに結び付けられた主題語は 102 件に留まった。また、その 102 件の主題語に、結びついたマンガの数は図 3 のようになった。

5.4 主題語彙の拡充

(1)マンガのレビュー文からの取得結果

マンガのレビュー文について形態素解析を行うことで単語を抽出した。形態素解析を行う前に、レビュー文中の絵文字に対して正規表現を用いることで排除した。また、カタカナの文字列について半角カナと全角カナが混在している場合が多く見られたため、半角カナを全角カナに変換する正規化を行った。レビュー文の形態素解析には MeCab を使用し、辞書には mecab-ipadic-neologd を採用した[8]。形態素解析を行う際に、主題語となりうる単語を抽出するために名詞の原型のみを抽出対象とし、さらに人名を排除した。 1 件以上のレビュー文に紐つけられた主題語 207 件に対し、それぞれ特徴量の高い順に特徴語を 50 件ずつ取得した。しかし、主題語ごとに結びつくレビュー文の数が大きく異なるため、多くの主題語に共通する単語の特徴量が高くなってしまった。そのため、10 件以上の主題語の特徴語として現れた

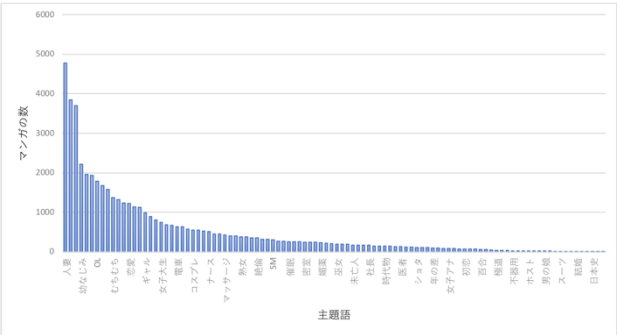

図 3 一件以上のマンガに結びつく主題語

語を排除した。その結果取得できた特徴語の中から、主題語の候補とする単語を人手で選択し、決定した。

本研究では、マンガとの結びつけ結果から、500 件以上のマンガと結びついた主題語 28 件を主題語彙における拡充が必要である箇所とし、そのうち 1 件以上のレビュー文に結びついた主題語 27 件について主題語の候補の取得を行った。その結果取得した主題語の候補の一部が表 5 である。主題語の候補として適当であると判断した単語は、27 件の主題語について全部で 116 件取得でき、 1 つの主題語あたりの平均は 4.3 件（小数点第 2 位を四捨五入）であった。

(2)外部語彙からの取得結果

オンライン百科事典の一つに、イラスト・マンガ・小説の投稿・閲覧に特化した Web サイトである pixiv[4] に投稿される作品に付与するタグを解説する記事を主に構成されるピクシブ百科事典がある[9]。これは、ピクシブ百科事典の記事ではその特性上、マンガの主題語に見られるような、単語のもつ暗黙的な意味や、それに基づく関連などが記述される傾向がある。そこで、本研究では情報を抽出する外部語彙として、ピクシブ百科事典を選択した。

主題語の候補の抽出は、ピクシブ百科事典における主題語の単語記事にアクセスし、そこに記述される関連単語の情報を、Web スクレイピングによって取得した。まず、主題語彙中の主題語についての記事がピクシブ百科事典に存在するかを調査する。これは、ピクシブ百科事典における記事の URL が、「https://dic.pixiv.net/a/(記事名)」となっていることを利用し、URL に「https://dic.pixiv.net/a/(主題語名)」としてアクセスしてすることで行った。そして、記事が存在する場合は、記事の中に記述される「親記事」「子記事」「兄弟記事」に記述される単語を関連単語として取得した。

主題語彙中の主題語についての記事をピクシブ百科事典から取得したところ、別名を含み重複を排除した主題語 958 件のうち、751 件の主題語について単語記事が存在した。そのうち、500 件以上のマンガと結びついた主題語 28 件について、主題語の候補の取得を行ったところ、28 件中 27 件の主題語についてピクシブ百科事典に単語記事が存在した。取得できた関連単語数について、最も多かった主題語では 20 件の関連単語が取得できた。また最も少ないものでは 2 件のみに留まった。また、27 件の主題語について平均で 12.8 件（小数点第 2 位を四捨五入）の関連記事が取得できた。

[4] https://www.pixiv.net/

表 5 レビュー文から取得した主題語の候補

主題語	主題語の候補とした単語
同級生	スポーツブラ、クラスメイト、友達、高校
人妻	ヨガ
調教	マスターベーション、名門校、上役、射精、学内、清純派、子供、パンチラ、メガネっ子、連れ子、日本人、階段
幼なじみ	パブ、鈍感
ファンタジー	あやかし、カエル、人魚、童話、麒麟、猪、神様、蛇、獣、鬼、異種、犬、蛙、人魚姫
OL	課長、師、会社、社員旅行、デスク

表 6 主題語の候補としなかった語

	レビューから取得した特徴語
マンガの主題を表現するものではない語	真っ先、さっき、間、それ、化、日、作り、急、周り、思い、なに、場面、ベスト 3、余地、構造、誰、いつ、前、理
キャラクター名と思われる語	かなた、愛実、あいちゃん、リリィ、lily、Lily、一之瀬
マンガの形態や販売に関する語	1 巻、完結、フルカラー、合本、一巻、短編集、読み切り、通常価格、巻、発売決定
感想に関する語	満足、控えめに言って最高、すき

6.考察

主題語彙をマンガと結びつけた結果、多くのマンガに結びつく主題語が存在することが明らかになった。また、マンガに結びつく主題語は、主題語彙の中の一部の主題語に留まり、その中でも一部の主題語にマンガが集中していることがわかる。これらの主題語が表すのは、多くのマンガに共通する主題であり、他のマンガとの主題の差異を明確にすることができない。そこに主題語を拡充することで、マンガの主題のより細かな差異を表現できる主題語彙を構築することができると考えられる。

5.4 で述べたように、レビュー文から取得できた主題語の候補の数は、1 つの主題語あたり数件程度であった。主題語彙とマンガの結びつけ結果に見られた大量のマンガに結びつく主題への主題語の拡充を考えると、十分な数が取得できているとはいえない。そのためより多くの語が取得できるように手法の改善を図る必要がある。また、取得できた主題語の候補となる語をみると、例えば主題語「同級生」について抽出した「クラスメイト」や「高校」といった単語が、主題語の表す概念をより細分化するために使用できる単語であるといえる。しかし、こういった単語が抽出できたのは少数であった。また、主題語「ファンタジー」について抽出した「人魚」や「カエル」などの単語は、主題語「ファンタジー」を付与されたマンガの主題の差異を、登場するキャラクターによって区別するために使用することができると考えられる。このように、主題語との間に意味的な関連を推測することが困難な単語であっても、物語の舞台や登場人物の身体的特徴などのマンガの主題を表現しうる単語であれば、主題語との関連を記述することでマンガの主題のより詳細な差異を表現することができる主題語彙を構築することができる。

レビューから抽出した特徴語について表 6 に示したような、マンガの主題を表現するものではない単語が多く抽出された。そこには、形態素解析を行う際に取り除くことができなかったキャラクターの名前と思われる単語や、マンガの形態などに関する単語といったものが見られた。そのため、レビュー文から多数の特徴語が取得できているにも関わらず、主題語の候補となるような語がほとんど取得できない主題語が見られた。こういった語を排除することで、主題語ごとに特徴的な単語を取得することができ、より多くの主題語の候補となる語が取得できることが考えられる。また、本研究では、主題語ごとにレビュー文の数が大きく異なったため、単語の特徴量を算出する手法として TF-IDF が適当でなかった可能性がある。今後は、主題を表現するとは言えない語を排除するためのストップワードの辞書の作成や、単語の特徴量の算出方法およびレビュー文を収集する情報源の検討を行う必要があると考えられる。

ピクシブ百科事典に記述されている主題語の単語記事に記述されている各関連記事に記述されている

語を調査したところ、「親記事」には主題語に対して広義の意味をもつ単語が、「子記事」には主題語の表記違いの単語あるいは主題語に対して狭義の意味を持つ単語が、「兄弟記事」には主題語と何らかの上位概念を共有していると考えられる単語が記述されている傾向があることがそれぞれ明らかになった。

ピクシブ百科事典から取得した単語を用いて、最も多くのマンガに付与されていることがわかった主題語「人妻」について主題語の拡充を行った例を示す。主題「人妻」に関する主題語の候補を取得するために、ピクシブ百科事典の「人妻」の単語記事における親記事・子記事・兄弟記事から抽出した関連単語を記事中に記述されていた関連に基づいて主題語彙に当てはめたものが図 4 である。ただし、親記事の単語の下位語として兄弟記事の単語を記述し、既存の主題語と重複するものは排除した。この図では右の楕円が左の楕円の下位の主題であることを表現しており、例えば「花嫁」は「人妻」の下位の主題である。また白の楕円があらかじめ主題語彙に存在していた語、青が親記事から取得した語、黄が兄弟記事から取得した語、赤が子記事から取得した語をそれぞれ表現している。この結果から、一部主題語の関係を人手で修正する必要があると考えられる箇所が見られるが、主題語をより詳細な概念に分ける単語や類似する単語など、主題語彙を拡充するための主題語の候補が取得できているといえる。よって、マンガの主題語彙を拡充する主題語の候補を、外部語彙であるピクシブ百科事典に記述された関連単語から取得することが有効であると考えられる。

しかし、記事から取得した関連単語には「巨乳の人物一覧」や「淫乱カグラ」のような、pixiv のイラストのカテゴリを表現するためのものや、特定の作品などを元ネタとするものなど、pixiv のコミュニティの中で独自の使われ方をするものが見られた。こういった語はマンガの主題語としては適切であるとは言えないため排除する必要がある。また、主題語ごとに取得できる関連単語の数が異なり、主題語によっては取得できる主題語の候補の数が少なくなる場合が見られた。そういった主題語については、複数の外部語彙から情報の抽出を行い取得できる関連単語の数を増やすなど、その改善を行う必要がある。

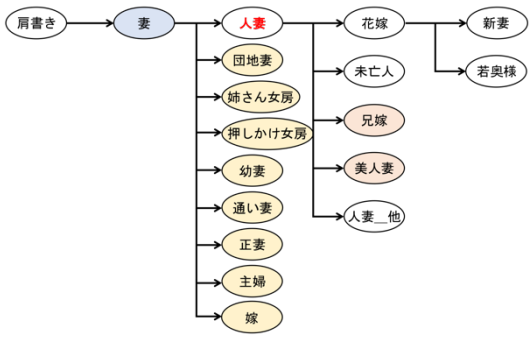

図 4 外部語彙から取得した主題語の候補

7.おわりに

本研究では、電位書籍ストアで提供されるマンガの書誌情報と、人手によって構築されたマンガの主題語彙のリンキングによって主題語彙の拡充を行い、主題検索に適した主題語彙の構築手法を提案した。電子書籍ストアのタグと主題語彙をリンキングすることで、多くのマンガに付与されている主題語を発見し、主題語彙における主題語の拡充が必要な主題を発見することができた。また、マンガのレビュー文と既存の外部語彙を利用することで、主題語彙を拡充するための主題語の候補を取得することができた。

今後の課題として、取得できる主題語の候補の数が少ない問題や、主題語の候補として適当でないものを取得してしまう問題などに対して、レビュー文に形態素解析を行う場合のストップワードの辞書の作成、取得した単語の特徴量算出の手法の改善、レビュー文を取得する情報源あるいは情報を抽出する外部語彙の検討などを行うことが考えられる。

謝辞

本研究は JSPS 科研費 JP18K11984 の助成を受けたものです。

参考文献

[1] 2018 年版出版指標年報. 公益社団法人全国出版協会・出版科学研究所, 2018, 389p.

[2] もり・きよし. 日本十進分類法. 新訂 10 版, 公益社団法人日本図書館協会, 2015.

[3] 表智之,金澤韻,村田麻里子. ”マンガとミュージアムが出会うとき”. 臨川書店, 2009, 259p.

[4] 馬場こづえ, 藤井敦, 石川徹也. “小説テキストを対象としたジャンル推定と人物抽出”. 言語処理学会第 11 回 年次大会予稿集. 2005, p. 157−160.

[5] 山下諒,松下光範.” 階層的トピック分類を用いた内容情報に基づくコミック探索システムの提案”.HCG シンポジウム.2015.

[6] David. M. Blei, Thomas. L. Griffiths, and Michael. I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. Journal of the ACM, Vol. 57(2), No. 7, 2010.

[7] “SKOS Simple Knowledge Organization System Refe

rence”. W3C. https://www.w3.org/TR/2009/REC-skos-reference-20090818/, (参照 2019-2-3)
[８] “neologd/mecab-ipadic-neologd”. mecab-ipadic-NEologd : Neologism dictionary for MeCab. https://github.com/neologd/mecab-ipadic-neologd/, (参照 2019-1-24)
[９] ピクシブ百科事典製作委員会. ”ピクシブ百科事典”. https://dic.pixiv.net/, (参照 2019-2-3)