坪田譲治の児童文学作品におけるエピソードの検索

川上 隆人† 劉 渤江‡ 北川 文夫‡

†岡山理科大学大学院総合情報学部情報科学専攻

†‡岡山理科大学総合情報学部

E-mail: †stanley0305@hotmail.com ‡{liu, kitagawa}@mis.ous.ac.jp

あらまし 岡山市では，児童文学作家である坪田譲治の文学データベースを構築している．文学研究者から特定のエピソードを作品から検索する要望があった.エピソードはストーリ性のあるセンテンスのリストとして考えられ，多数の作品に如何に漏れなく速く特定できるかが課題となる．本稿では，作品から動作と場面をあらわす語句に着目することで，エピソードを特定する方法を提案する．そして，検索アルゴリズムの実現方法について述べる．

キーワード 情報探索，テキストDB，インデックス

Search of the episode in Jouji Tsubota's juvenile literature work

Takato KAWAKAMI † Bojiang LIU ‡ and Fumio KITAGAWA ‡

†Okayama University of Science synthesis information faculty information science speciality

‡Okayama University of Science synthesis information faculty

†‡1-1, Ridai-cho, Okayama-shi, 700-0005

E-mail: †stanley0305@hotmail.com ‡{liu, kitagawa}@mis.ous.ac.jp

Abstract In Okayama-shi, the literary database of Joji Tsubota who is a juvenile literature writer is built.From a literary researcher, there was a request which searches specific episode from a work.It becomes a subject how episode is considered as a list of sentences with story nature, does not leak to many works, and can be specified quickly. In this paper, the method of specifying episode is proposed by paying one's attention to the words and phrases which express operation and a scene from a work.And the realization method of search algorithm is described.

Keyword information search, Text DB, an index

1. はじめに

岡山市では，児童文学者の坪田譲治に関する文学研究データベース[1]の構築を行っている．このデータベースは坪田譲治の作品や研究資料など，様々な情報を一元的に管理するものであり，近代児童文学の研究へと多いに貢献すると期待されている．構築にあたり，データの検索方法に関して文学研究者からいくつかの要望があった．それらの要望は文学作品の書誌情報などによる一般的な作品の検索方法ではなく，坪田譲治の作品に見られる独自な特徴に基づく検索方法であった．その1つの検索方法にエピソードによる作品の検索が挙げられた．

小説などの文章から読みたい一部分を探し出すのは手間のかかる作業である．また，複数の作品や読んだことのない作品などから，特定の一部分を探すには多大な手間と時間を必要とする．児童文学研究者が研究を進めるにあたり，「複数の作品で同じ内容のエピソードが登場する」といった，坪田譲治の独自な文体が問題として挙げられた．実例に挙げられたエピソードに，「米倉に穴があいて，米が盗まれる」といった出来事がある．このエピソードは，坪田譲治の作品である「人見のおばあさん」[2]，「どろぼう」の双方に類似した文章が掲載されている．このような，与えられたエピソードから，その内容が実際に記載されている作品を検索する方法が，「エピソードの検索」である．

文章から特定の情報を検索する手法として，TF×IDF[3][4]などの検索手法が存在する．しかし，この手法では文章の意味的な解釈が弱く，エピソードを特定するには精度が十分であるとは言えない．エピソードの検索を行うためには,内容を意味的に解釈し,また,

作品のどの部分に記載されているかを特定する必要がある．

そこで，本研究ではエピソードの検索を実現する手法として，作品のパラグラフごとに各センテンスから主語，述語，目的語のメタデータを抽出し，それらのメタデータの類似性から目的のエピソードを検索するアルゴリズムを提案する．そして，実装することで本手法の有効性を示す．

まず，2 節で文学研究者の要望をいくつか紹介する．また，3 節でエピソードの検索方法について説明し，4 節で実現方法と検索速度を向上するためのインデックスについて述べる．そして 5 節では，実際に検索を行ない，検索結果について考察を述べる．

2. 坪田譲治研究データベース

ここでは，坪田譲治研究データベースの構築にあたり児童文学研究者から要望があった検索方法をいくつか紹介する．これらの検索方法はいずれも，坪田譲治の作品にあらわれる特徴によるデータの検索方法である．それは，以下のような検索である．

（1）エピソードによる作品の検索
（2）作品の類似性による作品の検索
（3）登場人物による作品の検索
（4）方言による作品の検索

2.1. エピソードによる作品の検索

坪田譲治の作品では，同じ内容のエピソードが複数の作品に登場することがある．例えば，「米倉に穴があいて，米が盗まれる．」といったエピソードは，「人見のおばあさん」と「どろぼう」の双方の作品で登場している．このような，あるエピソードを与えることで，そのエピソードが記述されている作品を提示する検索である．

2.2. 作品の類似性による作品の検索

作品には，編集の遂行により類似性の確認できる作品が存在する．例えば，「妹とカタツムリ」は部分的な修正を加えられることにより，題名を変えて「カタツムリ」という作品として発表されている．また，この「カタツムリ」も修正を加えられることにより「でんでん虫」として発表されている．

部分的な編集も存在すれば，ある作品の一部分を切り抜いて，別の作品として発表している編集も存在する．このような編集による作品の類似性により，作品を検索する方法である．

2.3. 登場人物による作品の検索

坪田譲治の作品では，複数の作品で同じ名前の人物が登場することがある．しかし，これら人物は異なる設定を与えられており，まったく別の人物として作品に登場している．例えば「正太，善太，三平」の 3 人は多くの作品に共通して登場するが，作品別に，親族，友人，故人など異なる人間関係を与えられている．このような登場人物の人間関係により，作品の検索を行う方法である．

2.4. 方言による作品の検索

坪田譲治の出身地は岡山県であり，岡山の方言が使われている作品がいくつか存在する．しかし，編集の段階で別の言葉に置き換えられた方言や，作品中で意図的に書かれた方言など，坪田譲治の執筆環境により方言の使われ方はそれぞれ異なっている．このような方言が記述されている作品を提示する検索方法が挙げられた．

3. エピソードの検索

この節では，児童文学研究者によって挙げられた検索方法の一つである「エピソードの検索」について説明する．実例として挙げられた作品を読んで，エピソードについて考察を行い，具体例を用いてその検索方法について述べる．

3.1. エピソードの定義

エピソードとは本筋の間にはさむ，本筋とは直接関係のない短い話である．しかし，児童文学研究者が実例として挙げたエピソードは，本来のエピソードの意味とは少し異なり，作品内で述べられている「ある出来事」を表していた．

そこで，エピソードとは「作品の本筋との関係の有無に限らず，作品内で述べられている出来事」と定義する．以下に児童文学研究者が挙げた，いくつかの実例を列挙する．

- 米倉に穴があいて、米が盗まれた
- 小学校でイタチを捕まえて遊ぶ
- 川蟹をとるための網を作る
- 子供が本を読んでいて分からない字がありおとうさんに読み方を聞く

3.2. エピソードの範囲

作品内の文章に対して，エピソードがどこから始まり，どこで終わるのかを明確にしなければエピソードは特定できない．実例として挙げたエピソードが掲載されている作品では，エピソードは一文であったり，複数の文であったりその範囲は様々である．しかし，いずれのエピソードも，「物語のある一場面に含まれる」ことが作品から読み取れる．

坪田譲治の作品は，ストーリ性のある物語である．物語はいくつかの場面によって構成されている．場面とは，物事が行われているその場の様子である．場面は作品内での「時間，場所，人」などの情報の変化により構成されると考えられる．具体的に作品を確認したところ，これらの情報はパラグラフによって変化す

ることが多いことが確認できた．そこで，本研究ではパラグラフをエピソードの単位と考えることで，その範囲をある程度，収束できると考えた．

3.3. エピソードの判断基準

複数の作品で同じ内容のエピソードが出現する問題において，何を判断基準としてエピソードが同じ内容であるか明確にする必要がある.そこで,「単語の類似性」と,「単語の関連情報」に着目する．実例として挙げられた,「米倉に穴があいて，米が盗まれた」のエピソードは,「米倉，穴，開く，米，盗む」といった単語で構成されている．実際に，このエピソードが掲載されている作品は「人見のおばあさん」(図 1),「どろぼう」(図 2)であるが，これらの作品に用いられている単語を見ると，エピソードで上げられた単語と類似性を確認することができる．また，このエピソードを読むと文法から,「米倉，穴，開く」,「米，盗む」のような単語の組み合わせを認識することができる．これら類似性の確認できた単語の組み合わせも，エピソードが掲載されている作品と類似していることが確認できる．そこで，単語の意味による類似性と，他の単語との関連情報を，エピソードを識別する判断基準と考えた．

さて、そのあくる朝です。おばあさんが戸をあけてみると、米倉には大穴があいていました。そしてお米は二十俵ほどぬすまれていました。やはり、船できて、倉のかべをやぶって、そこから運び出したものです。

図 1 「人見のおばあさん」のパラグラフ 23

昨夜の間に、どろぼうは米倉の外側を流れている川に一艇の船を引いて来て、倉の壁を切り破り、そこから五十俵もの米を盗んでしまったのでした。さて、その夜のことお爺さんの子に、松山お米、即ち善太のおばあさんが生れました。お米をとられたというので、こんな名をつけたのだそうであります。

図 2 「どろろぼう」のパラグラフ 14

3.4. エピソードの表現方法

本稿では，エピソードとは「作品の本筋との関係の有無に限らず，作品内で述べられている出来事」と定義した．そこで，本研究では，検索するエピソードの出来事と，作品内で記述されている出来事を「主語・述語・目的語」の三次元データの組み合わせで表現することにより，エピソードの検索を実現する方法を提案する．これにより検索するエピソードで述べられている出来事と，作品内に記載されている出来事の全ては，三次元データのリストとして表現される．この三次元データを照合することで，エピソードの検索が可能になると考えられる．

3.5. エピソードの照合

検索エピソードから抽出した三次元データと，検索の対象とする作品の三次元データを，パラグラフを単位として比較する.検索エピソードの「主語(A)・述語・目的語(B)」の組み合わせと，この三次元データの主語と目的語を入れ替えた「主語(B)・述語・目的語(A)」の組み合わせで,以下のパターンにて照合は行われる.

1. 「主語(A)，述語，目的語(B)」で照合
2. 「主語(A)，述語」で照合
3. 「述語，目的語(B)」で照合
4. 「主語(B)，述語，目的語(A)」で照合
5. 「主語(B)，述語」で照合
6. 「目的語(A)，述語」で照合

主語と目的語を入れ替えデータで照合を行うのは，受身の状態で表現されたセンテンスを想定してのことである．また照合の際，単語の意味的な距離も考慮に入れて行う．例えば,「米倉」と「倉」は共に「家屋」を表している語句であり，単語の意味的には距離が近いと考えられる．照合の際に，これらの作業を行うことにより，エピソードの検出漏れを減らすことが可能になると考えられる．

3.6. 具体例

具体例に「米倉に穴があいて，米が盗まれた」のエピソード用いて，エピソードの照合方法について説明する．

まず，与えられた検索エピソードである「米倉に穴があいて，米が盗まれた」のセンテンスから主語，述語，目的語の三次元データを抽出する．このエピソードは動詞として「あく」「盗む」の 2 つの単語を持っている．そのため，2 組の三次元データ(表 1)が抽出される．抽出において本研究では，目的語をセンテンスの動詞に関連のある名詞とする．これにより，目的語は一般的に考えられている目的語より広義な意味で扱われ，形容詞的な働きをする単語なども目的語として抽出する．また，このエピソードでは「米が盗まれる」の目的語の該当語が存在しない．このように該当語が存在しない場合は「NULL」と表現する．

表 1 検索エピソードの三次元データ

主語	述語	目的語
穴	あく	米倉
米	盗む	NULL

この検索エピソードは「人見のおばあさん(パラグラフ 23)」(図 1)と,「どろぼう(パラグラフ 14)」(図 2)の双方の作品にて掲載されている.「人見のおばあさん(パラグラフ 23)」と「どろぼう(パラグラフ 14)」の三次元データは，次のように表される．(表 2)(表 3)

表 2 「人見のおばあさん(パラグラフ 23)」

主語	述語	目的語
おばあさん	あける	戸
穴	あく	米倉
米	盗む	NULL
NULL	くる	船
NULL	破る	壁
NULL	破る	倉

表 3 「どろぼう(パラグラフ 14)」

主語	述語	目的語
どろぼう	くる	NULL
川	流れる	外側
川	流れる	米倉
どろぼう	引く	船
どろぼう	きる	倉
どろぼう	きる	壁
どろぼう	破る	倉
どろぼう	破る	壁
どろぼう	盗む	米
おばあさん	生まれる	子
NULL	とる	米
NULL	つける	名

「人見のおばあさん(パラグラフ 23)」の三次元データのリストでは,「穴，あく，米倉」,「米，盗む」の組み合わせが，検索エピソードと一致していることが確認できる.

「どろぼう(パラグラフ 14)」の三次元データのリストでは,「盗む, 米」が検索エピソードと一致していることが確認できる. また,「破る, 倉」は「あく，米倉」と単語に類似性があり,「とる，米」は「盗む，米」も類似性があることが確認できる.

これらの組み合わせを照合することにより，エピソードの検索は可能となる.

4. 実現方式

この節では，エピソードの検索を実現するにあたっての処理の流れについて説明する. エピソードの検索は以下の手順で実現される.

1. 文章の構造化
2. 主語・述語・目的語の抽出
3. インデックスの作成
4. 検索アルゴリズム

4.1. 文章の構造化

エピソードの範囲は，パラグラフを単位とすることである程度に収束できると述べた. そこでまず作品からパラグラフの抽出を行なう. そして，パラグラフからセンテンスを割り出し,文章を階層構造で表現する. このパラグラフとセンテンスには一意な ID を付与する. ID は，パラグラフ，センテンスの位置を特定できる識別子である.

4.2. 主語, 述語, 目的語の抽出

センテンスから主語，述語，目的語のメタデータを抽出する. 抽出はセンテンスに記述されている単語に関して,文法的な係り受けを調べることで実現される.

抽出には，まずセンテンスから述語を特定し，その述語に受け係りのある単語の中から主語，目的語を特定する. 述語として抽出する単語は，センテンスに記述されている動詞で,「自立動詞」である単語とする. 主語として抽出する単語は，述語に係り受けのある単語で，助詞「は」,「が」を伴う「名詞」とする. また，目的語として抽出する単語は，述語に係り受けのある単語で,助詞「の」,「を」,「に」,「へ」,「と」,「より」,「から」,「で」を伴う「名詞」とする.

抽出した単語には辞書を用いて意味付けを行なう. 辞書には日本語語彙大系[3]を用いることにした.

以上の操作により，全てのセンテンスは,「主語，述語，目的語」の三次元データのリストとして表現され，三次元データの掲載位置を示す ID と共にデータベースへ格納される.

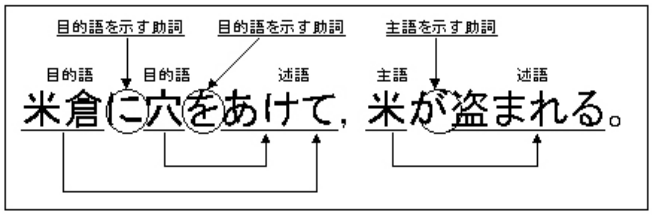

図 3 係り受けによる主語・述語・目的語の特定

4.3. インデックスの作成

作成された三次元データのリストは，膨大な数のデータを持っている. そのため，リストから一つひとつ照合するは能率的ではない. そこで，まず，エピソードが含まれる可能性のあるパラグラフをいくつかに絞り込み，絞り込まれたパラグラフの中からより詳細な三次元データの照合を行なう. これにより検索速度の向上を図る.

まず，各作品のセンテンスから抽出したメタデータを，主語，述語，目的語によりソーティングを行う. 次に，ソーティングされた主語，述語，目的語を意味別に分類する. この分類の項目は，日本語語彙大系の「一般名詞意味属性体系」,「固有名詞意味属性大系」,「用言意味属性体系」のカテゴリーに基づいた分類とする.日本語語彙大系は単語を 3000 カテゴリーで分類し，階層構造によって表現している. この階層構造では名詞を最深部 12 の階層,動詞を最深部 4 の階層で表現している. 以上のような分類によりインデックスの作成を行う. (図 4 に参照)

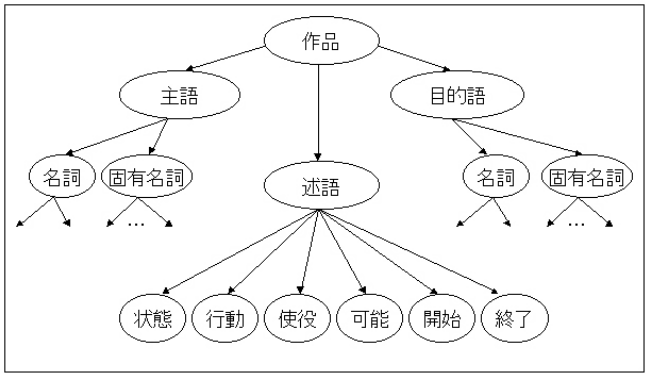

図 4 インデックスによる語句の階層構造

4.4. 検索アルゴリズム

検索は以下の流れにて実現される.

（1） 検索エピソードとして与えられた文章から主語, 述語, 目的語の三次元データを抽出する. そして, インデックスを参照するため, それぞれの単語を日本語語彙大系によりの分類わけを行う.

（2） 分類した単語を各作品の主語, 述語, 目的語と比較する. 例えば, 主語が「おばあさん」であるなら, インデックスで作成した階層構造「作品→主語→名詞→…→人」のリストへ移動し,「人」の項目にて照合を行う. また, この三次元データの主語と目的語を入れ替えて, 再度検索を行う.

（3） 候補となる主語のパラグラフ ID, 述語のパラグラフ ID, 目的語のパラグラフ ID を割り出す. そして, それぞれの ID の積集合を計算する.

（4） 検出されたパラグラフは, 主語、述語、目的語の組み合わせをパラグラフ単位で照合しているためエピソードの有無が曖昧である. そこで再度,「主語, 述語, 目的語」の組み合わせをセンテンス単位で照合して, 最もエピソードと類似性の高いパラグラフを特定する.

5. 実験

この節では, エピソードの検索結果について考察する. 検索システムの実装にあたり, プログラム言語には Java, データベースには MYSQL を利用した. また, 係り受けの解析には日本語構文解析器である CaboCha[6]を用いた.

実験として, 児童文学研究者によって挙げられたエピソードをいくつか検索する. これらのエピソードはあらかじめ掲載されている作品が判明しているものである. 検索結果において, 作品を正しく検出できるかを試す.

5.1. 検索結果

（1） 米倉に穴があいて, 米が盗まれる

このエピソードからは,「穴, あく, 米倉」,「米, 盗む, NULL」の 2 組の三次元データが抽出される. 検索では以下の 2 つの作品が検出された.

- 人見のおばあさん(パラグラフ 23)
- どろぼう(パラグラフ 14)

これらの作品は, あらかじめ掲載作品が判明している作品と一致している. 人見のおばあさんでは, 2 組の三次元データが共に適合し, どろぼうでは,「米, 盗む, NULL」の三次元データが適合していた.

（2） 小学校でイタチを捕まえて遊ぶ

このエピソードでは, 該当する作品を検出できなかった. 検出に失敗した原因は, 掲載作品では,「イタチを捕まえる」ではなく「イタチをつりあげる」と表記していたからである.

（3） 川蟹をとるための網を作る

このエピソードでは, 検索結果として以下の 3 つの作品が検出された.

- 木の葉がに(パラグラフ 20, 21, 22, 27, 30, 42, 51, 63)
- イタチのいる学校(パラグラフ 18)
- 川干と胴じり網(36)

実際に, このエピソードが記載されている作品は「木の葉がに」である.「イタチのいる学校」,「川干と胴じり網」の作品では「網を作る」といったエピソードが記載されていることで検出されたノイズである. また,「木の葉がに」の作品では,「蟹をとる」,「網を作る」といった出来事が, 異なるパラグラフで述べられていることにより 8 つのパラグラフが検出された.

（4） 子供が本を読んでいて分からない字がありおとうさんに読み方を聞く

このエピソードでは, 次の作品が検出された.

- おとうさんの話(パラグラフ 2・10・11・12)
- かっぱに出会った話し(パラグラフ 0)
- コオロギの話し(パラグラフ 2)

このエピソードでは,「NULL, 読む, 本」,「子供, わかる, NULL」,「字, 聞く, おとうさん」の 3 つの三次元データが抽出でされた. 実際にこのエピソードが登場しているのは「おとうさんの話」であるが, 3 つの三次元データが, 1 つのパラグラフで述べられるのではなく, 複数のパラグラフに分散して述

べられていた．その他の検出された作品，「かっぱに出会ったはなし」,「コオロギのはなし」は「NULL，読む，本」の三次元データによって照合されたノイズであった．

5.2. 考察

いくつかのエピソードでは，掲載された作品とそのパラグラフを特定することに成功した．しかし，検出できなかったエピソードや，パラグラフの位置を特定するには至らなかったエピソードも存在した．

検出に失敗した理由の一つに，「単語の類似性が不明瞭」であることが考えられる．これは，特に動詞として用いられている単語に原因がある．例えば，「米倉に穴をあけて，米を盗む」のエピソードと，「米倉を破って，米を盗む」のエピソードがあるとする．これらエピソードの，「あける」と「破る」の動詞の単語から，センテンスの意味的な類似性を見出すことは現状では困難である．

エピソードが掲載されている作品は特定できたものの，パラグラフの特定に失敗した理由として，次のようなことが考えられる．検索するエピソードが作品に記載されていても，複数のパラグラフにまたがっている場合は検出できない．この問題は，物語の一場面が，複数のパラグラフで構成される場合があることに発起する．そのため，場面の区切りを一つパラグラフではなく，複数のパラグラフで表現するなどの対処が必要である．

また，照合の際に受身で表現されたセンテンスを想定して，主語と目的語を入れ替えた三次元データで照合を行った．しかし，この操作ではセンテンスの意味とは異なった内容で照合を行ってしまう場合が考えられる．

6. まとめ

本研究では，児童文学作家である坪田譲治の文学研究データベースのデータ検索方法として，児童文学研究者の視点に基づいた検索方法である「エピソードの検索」の実現方法について述べた．方法として，エピソードの範囲を収束するためにパラグラフを単位として，使用されている単語の類似性と単語の関連性についてセンテンスから，「主語・述語・目的語」を抽出することでエピソードを表現した．そして，実際に検索システムを作成することにより，坪田譲治の作品の中から，エピソードが含まれている可能性のあるパラグラフをいくつか特定することに成功した．今後の課題としては，以下のような問題が挙げられる．

- 代名詞を考慮した三次元データの生成
- 主語の省略されたセンテンスから主語の抽出
- 記載されている出来事の時間軸を考慮した三次元データの抽出
- 本手法と，他の手法との有効性の比較

これらの問題が今後の課題である．また，児童文学研究者の要望として挙げられた他の検索方法についても，実現していく必要があるだろう．

謝　辞

本研究を進めるにあたり，多くの支援を頂いたノートルダム清心女子大学文学部日本語文学科助教授山根知子助教授，岡山市企画局総合政策部文化政策課ならびに坪田譲治研究データベース作成委員会の皆様に深く感謝申し上げます．

文　献

[1] 坪田譲治研究データベース，坪田譲治研究 DB 作成委員会，http://crane.mis.ous.ac.jp/.

[2] 坪田譲治，坪田譲治全集第 10 巻，佐藤亮一，新潮社，東京都，1978.

[3] 白井諭,大山芳史,池原悟,宮崎正弘,横尾昭男,“日本語語彙大系について”, 情報処理学会研究報告「情報メディア」アブストラクト,No.034-009,Nov 1998.

[4] Salton,G. and Buckley, C.: Term-weighting approaches in automatic text retrieval, *Information Processing and Management*, Vol.24, pp.513-523(1988).

[5] Salton,G. and Buckley, C.: Improving retrieval performance by relevance feedback, *Journal of the American Society for Information Science*, Vol.41, No.4, pp.288-297(1990).

[6] 日本語係り受け解析システム「CaboCha/南瓜」，工藤 拓,松本 裕治, http://chasen.org/~taku/software/cabocha/.