4F1-2in2

レビューを用いたコミックの特徴抽出における固有表現の影響に関する調査

Investigating Influences of the Named Entities Extracted from Comic Reviews

朴 炳宣 *1　　松下 光範 *2
Byeongseon Park　　Mitsunori Matsushita

*1*2関西大学 総合情報学部
Kansai University, Faculty of Informatics

The purpose of this research is to develop a method to access comics based on their content information. We have been examined to extract the set of feature words by taking the characteristics of description in the comic reviews into account. However, the extracted feature words often contain words that hinder the user from guessing the contents of the comic. Especially, the named entities defined in the comic tend to cause the problem; the amount of information obtained by the user varies greatly along with the user ' s knowledge about the named entities. In this paper, we investigated the influence of the named entity extracted from the comic reviews as the feature words. Our experiment revealed that the user ' s understandability was improved when the proportion of the named entities was decreased.

1. はじめに

書籍販売サイトや Wikipedia でコミックの検索を行う場合，作品のタイトルや著者，掲載誌などの書誌に関する情報に基づいた詳細な検索が可能である．一方，作品の内容に関する情報をトリガとした検索手段は，ジャンル情報（e.g.,“ファンタジー”，“恋愛”）を用いた大まかな検索に留まっている．

このような課題を解消するため，山下ら [4] はコミックの内容情報に基づく検索システムの実現を目的として研究を進めている．山下らの研究ではコミックの内容情報を抽出する試みとして，コミックのレビューから記述的特徴に基づきレビューの特徴語群の抽出を行い，それらをコミックの検索時にユーザに提示することで検索に用いる判断材料となるようにしている（図 1）．このシステムを用いることで，ユーザは自らの嗜好に合致した作品にアクセスができるようになった．しかし，山下らの手法により抽出された特徴語の中には，作品名や登場人物，世界観に関する用語といった作品内で定義がなされた未知語に基づく表現（以下，固有表現と記す）が多く含まれている．コミックレビューに登場する固有表現はユーザが持つ作品に対する認知度によって内容推測に役立つ情報の量が大きく変化する．円滑なコミック探索のためには，ユーザが知らないコミックであってもそのコミックの特徴を捉えられる情報を提示することが望ましい．

本稿では，コミックレビュー文中に存在する固有表現が Term Frequency - Inverse Document Frequency (TF-IDF) 法や hierarchical Latent Dirichlet Allocation (hLDA) 法 [1] によって抽出されたレビューの特徴語群に及ぼす影響について調査を行った．さらに，レビュー中の固有表現を特定するために，Wikipedia とレビュー文からコーパスを作成し，Conditional Random Fields [2] を用いた機械学習行うことで固有表現抽出器を作成した．

2. 先行研究

山下ら [4] は，レビューに含まれている感想や意見，あらすじなど様々な情報を網羅的に抽出するための試みとして，レビュー文中に含まれる名詞と形容詞に着目した．形態素解析には日本語の形態素解析器である MeCab*1 を用いている．この時，レビュー文中の固有表現は内容を把握する上で参考になり得ない情報であると判断し，このような品詞に関してはあらかじめ分析対象から除外するとした．山下らはこの方針に基づき，MeCab による解析結果から名詞と形容詞以外の品詞で分類されている単語を除外し，各コミックの特徴語群として用いた．しかし，解析結果には固有表現が一般名詞として分類されていることが確認された [4]．これは形態素解析に用いられた MeCab では，コミックレビュー中に存在する固有表現を全て網羅した解析ができなかったことが原因である．

図 1: 山下らのコミック検索システム [4]

図 2 は山下らの手法によりコミックレビューから抽出したコミックの特徴を表す語群（以下，特徴語と記す）の一例を示す．特徴語の中には “海賊団” や “仲間” といった，作品を知らないユーザであっても単語の意味を把握できる一般名詞が含まれている．これにより，ユーザは検索時に特徴語に基づき各コミックの内容を推測できる．一方で，“ルフィ” や “ウソップ” などのキャラクター名や “アラバスタ” や “覇気” などの世界観に関する用語といった固有表現が含まれていることも伺え

連絡先: 朴 炳宣, 関西大学総合情報学部, 大阪府高槻市霊仙寺町 2-1-1, k281401@kansai-u.ac.jp

*1 http://taku910.github.io/mecab/ (2017 年 2 月 28 日存在確認)

図 2: 提示する特徴語群に含まれている固有表現の例

る．意味を把握できないことによりコミックとの関連を読み取れない単語はユーザが検索を行う際，円滑な情報の把握への妨げとなり得る．

こういった問題を解決するには，形態素解析の際に用いられる辞書にあらかじめ固有表現を登録しておくことが望ましい．コミックの固有表現を収集できる情報源として Web 上の多言語百科事典である Wikipedia [*2] などが挙げられるが，現状では本システムの対象となっている全てのコミックに対応していないため， Wikipedia のみを用いた解決は困難である．また，コミックレビューはユーザによりインターネット上で自由に作成されるため，ユーザによって用語の使い方が異なったり決まった形式を持たなかったりする．このような特性を持つコミックやコミックレビューに対して，人手で作成した抽出規則や辞書を用いることによる固有表現の抽出は困難である．以上のことから，レビューを用いた分析を行うためには，レビュー文中に存在している固有表現を自動的に抽出する手法が必要である．

本稿では，コミックのレビューに含まれている多様な固有表現を抽出するために，機械学習を用いることにより，抽出規則を動的に作成する．また，機械学習に用いるコーパスの作成には，コミックのレビューに登場している固有表現が定義された Web 上の情報とレビュー本文を組み合わせて用いることで，大量のコーパスを効率的に確保し，より多様な固有表現に対応できるコーパスを用いた学習を行う．

3. コミックレビューの固有表現抽出

3.1 Conditional Random Fields

本研究では機械学習のアルゴリズムとして，Conditional Random Fields (CRF)[2] を用いた．CRF は系列ラベリングのための識別モデルであり，形態素解析や固有表現抽出などに多く使われている手法である．CRF では

$$D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(|D|)}, y^{(|D|)})\} \quad (1)$$

のようなデータが与えられた時、その条件付き確率 $P(y|x)$ は

$$P(y|x) = \frac{1}{Z_{x,w}} exp(\omega \cdot \phi(x, y)) \quad (2)$$

$$Z_{x,w} = \sum_{y} exp(\omega \cdot \phi(x, y)) \quad (3)$$

で表される．

入力 x に対する最適な出力 y は，Veterbi アルゴリズムを用いて効率的に求めることができる．

*2 https://www.wikipedia.org/（2017 年 2 月 28 日存在確認）

3.2 固有表現クラス

本研究では，コーパスに付与する固有表現クラスとして，情報抽出・情報検索の評価型ワークショップである Information Retrieval and Extraction Exercise (IREX) で定義した固有表現クラス [*3] を用いた．IREX の日本語固有表現抽出タスクでは 8 種類の固有表現を定義し，それぞれの固有表現は重ならないとしている．しかし，関根ら [3] は，テキスト内の情報抽出や質問応答の適応分野の広がりを考えると，従来のような少数の分類では不十分であり，より多数な固有表現タイプを考慮する必要があると指摘し，固有表現タイプを 200 種類と大幅に増やした拡張固有表現を提案している．この手法は各分野に特化した固有表現抽出に多く用いられている．

そこで本稿では，コミックのレビューに存在する固有表現の中で，IREX で定義した既存の固有表現クラスとは異なる記述パターンを持つ，コミック特有の固有表現を円滑に抽出するために STORY，ITEM，SKILL の 3 つの拡張クラスを設けた．STORY は特定のエピソードやストーリを示す名称（eg.,“アラバスタ編” や “開戦の章”）を， ITEM は作品中に存在する道具や物の名称（eg.,“どこでもドア” や “ゴムゴムの実”）を，SKILL は少年漫画でよく見られるキャラクターが持つ技などの名称（eg.,“北斗百烈拳” や “かめはめ波”）を示している．

3.3 コーパスの作成

本研究では，機械学習に用いるコーパスとして，コミックのレビュー文と Wikipedia 上の情報を組み合わせたものを用いた．Wikipedia にはコミック内に登場するキャラクタ名や世界観に関する用語といった情報が多く含まれており，複数のユーザによってより細かく整理されていることが多い．そこで，本研究では Wikipedia に記載されている固有表現をコーパスを作成するための辞書として用いた．

本稿で機械学習に用いるコーパスを作成するために行った手順について述べる．まずレビューサイト上でも特にレビュー数が多いコミック 20 作品の Wikipedia ページに存在する単語を収集し，各コミックの固有表現辞書を作成した．この時，採用した固有表現のクラスに基づき，各単語ごとに固有表現クラスを指定している．次に，レビュー文を「漫画レビュー.com[*4]」と「作品データベース [*5]」から収集し，各文に対して形態素解析を適用した．最後に，各コミックの固有表現辞書に基づき各形態素ごとに素性とタグ付けを行うことで，機械学習に用いるコーパスを作成した．今回学習に用いるための素性は，表層，品詞細分類，文字種，現在の位置 i から $i \pm 1 \sim 2$ の位置にある形態素の表層と品詞，固有表現タグの 5 つである．

実装には Python（バージョン 3.5.1 ）を用いており，形態素解析には MeCab の Python 用ライブラリを，機械学習には，CRFsuite[*6] Python 用ライブラリ（バージョン 0.8.4）を用いた．実装した固有表現抽出器を評価した結果，適合率が 0.788，再現率が 0.458，F 値が 0.526 となった．

4. 実験

本章では，コミックレビューに対して TF-IDF や hLDA などの手法を用いてレビューから抽出した特徴語群やトピックにおいて，固有表現がユーザのコミックの内容推測へ及ぶ影響について調査を行う．

*3 http://nlp.cs.nyu.edu/irex/NE/df990214.txt（2017 年 2 月 28 日存在確認）
*4 http://www.manngareview.com/（2016 年 1 月 10 日確認）
*5 http://sakuhindb.com/（2016 年 1 月 10 日確認）
*6 http://www.chokkan.org/software/index.html.en/（2017 年 2 月 28 日存在確認）

4.1 実験目的

先行研究 [4] では TF-IDF 法を用いて抽出した各コミックの特徴語や hLDA 法を用いて生成したトピックに含まれた語の中には，ユーザがコミック検索システムを用いてコミックの検索を行う際に意味を把握できない語（e.g., 図 2 ）が含まれていたことよって，ユーザがコミックの内容を推測する際の負担となっていた．特に固有表現は，作品や作品に関連している固有表現の認知度によって情報量が大きく左右されるため，ユーザに提示する情報として適さないと考えられる．

一方で，先行研究で行った実験によれば，ユーザが知らないコミックの内容を推測する際に参考になった語として，有名作品の作品名（e.g., ドラゴンボール，ポケットモンスター）や有名キャラクターの名前（ドラえもん，ルフィ）があることが確認された [4]．つまり知名度の高い固有表現には「○○に似た作品」や「□□に似たキャラクター」といった，作品へのイメージをより具体的にする効果があると考えられる．そこで本稿では，固有表現がユーザに与える影響を把握するために，

条件（1） 既存の情報

条件（2） 固有表現抽出器により抽出した固有表現をすべて排除した情報

条件（3） 固有表現抽出器により抽出した固有表現の中で知名度の低い語のみ排除した情報

という 3 つの条件を設けた．

本実験における「知名度の低い語」は，レビューから TF-IDF 法を用いて算出された IDF 値が 6.0 以下のものと定めた．TF-IDF 法とは，文書データに含まれる単語の相対的な重要性を表す指標として広く用いられている分析手法である．TF-IDF 法における IDF 値とは，分析に用いられるすべての文書に出現している頻度を表す DF 値の逆数であり，IDF 値が高ければ高いほどその語は全文書において珍しい語となる．

4.2 コミックの特徴語の内容推測に関する実験

4.2.1 実験目的

コミック検索システムでは，TF-IDF 法により分析された各コミックの特徴語をユーザに提示することで，コミックの内容の推測が行なわれる．TF-IDF 法の特性上，レビューに存在する各コミックの固有表現は他作品のレビューでは出現しにくいため， TF-IDF 法の分析結果の上位に分類されやすい．そこで本実験では，異なった方法で固有表現を用いている 3 つの条件を比較することにより，コミックの内容を推測する際固有表現がユーザが及ぶ影響について調査を行う．

4.2.2 実験手順

実験協力者は，関西大学に所属している大学生 10 名（男性 4 名，女性 6 名）である．本実験の課題は，(1) 提示された語群（TF-IDF の分析結果）から予想されるコミックの内容の推定，(2) 推測した内容の根拠となる語の選択，である．また，実験協力者が円滑に課題に取り組めるように著者が例題を用いて回答して見せた．著者が例題を回答するにあたり，発話思考法を用いて実験協力者に思考内容を伝えながら課題の取り組み方を説明した．

内容を推定した結果は自由記述とし，単語（e.g., SF，未来），短文（e.g., リアルな SF，宇宙飛行士になるのが夢の主人公の物語），作品名（e.g., ワンピース，スラムダンク）などの制約を設けなかった．本実験で実験協力者に取り組んでもらう課題には，各条件の結果をそれぞれランダムに 10 コミックずつ選定した計 30 コミックの特徴語である．本実験では回答時間を設けず，すべての課題に回答した段階で実験終了とした．また，被験者には被験者の意思で実験を中止できることを伝え，これまでの説明に対する理解を確認した上で実験参加に対する同意を得た．

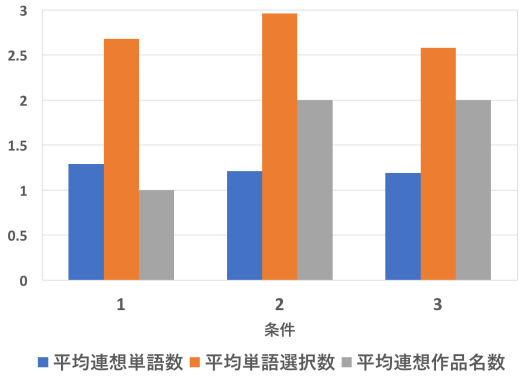

図 3: 各条件ごとの実験結果

4.2.3 実験結果

アンケート項目に対する実験協力者の回答結果の平均値を図 3 に示す．図 3 には，各条件に対するコミックの内容を推測した単語数の平均値，内容を推定する際に根拠となった語の選択語数の平均値，特定の作品名が連想できた数の平均値，の 3 つの数値が示されている．連想された単語数が最も多かったのは条件（1）であったが，選択した語数や連想した作品名数は条件（2）や条件（3）より劣っていた．実験参加者の回答を確認した結果，条件（1）では「シリアスホラー，少しだけコメディ」，条件（2）では「宇宙，コメディ」といった複数のジャンルを含めた回答が,「家族物語」,「野球漫画」といった一つのジャンルに絞られた回答が多いことがわかった．

4.3 コミックの特徴語とトピックとの関連性に関する実験

4.3.1 実験目的

コミック検索システムでは，各コミックの特徴語と各トピックに分類された語の共起に基づいて，コミックを関連づけている．このトピックを介したコミックの関連を正確に測ることができなければ，コミック検索時のユーザの思考とは異なる情報が提示されることになり，円滑な検索が難しくなる．そこで本実験では，異なった方法で固有表現を用いている 3 つの条件を比較することにより，システムで定義しているコミック間の関連（トピック）をユーザが推測する際に，固有表現が与える影響について確認する．

4.3.2 実験手順

実験協力者は，著者の所属研究室に所属している大学生，および大学院生 10 名（男性 4 名，女性 6 名）である．本実験では，まず実験協力者に対して実験目的と課題内容を説明した．本実験の課題の内容は，(1) 提示されたトピックと関連していると判断できるコミックの選択， (2) 提示トピックに含まれる話題と関連していると思われる各コミックの特徴語の選択，である．関連する語の選択時に制限を設けないことを伝えた．また，実験協力者が円滑に課題に取り組めるように著者が例題を用いて回答例を見せた．被験者には被験者の意思で実験を中止できることを伝え，これまでの説明に対する理解を確認した上で実験参加に対する同意を得た．

実験協力者に取り組んでもらう課題は，例題と同様に 1 つのトピックに対して 4 つの作品を提示しており，今回は各条件ごとに 3 つのトピックずつ，合計 9 トピックに回答しても

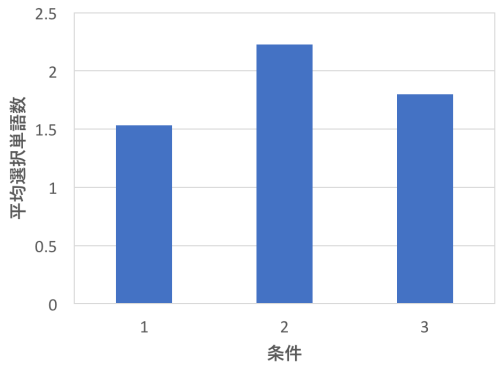

図 4: 各条件ごとの平均選択単語数

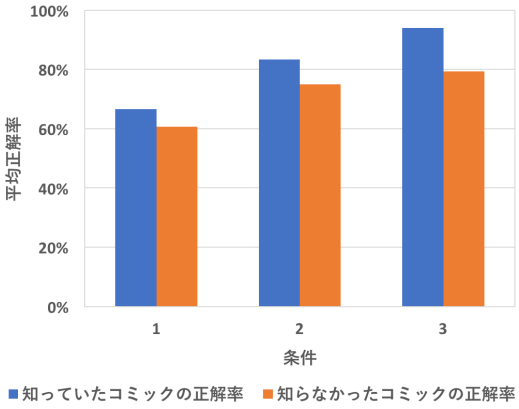

図 5: 各条件ごとの認知度による正解率

らった．また，課題に取り組んでもらった後に，回答を行うにあたり気づいた点があれば，自由記述で回答してもらった．最後に，実験に用いたコミックに関して，(1) 読んだことがある，(2) 読んだことがない，(3) 読んだことはないが知っている，の 3 択で回答してもらった．なお，本実験には回答時間を設けず，すべての課題に取り組んだ時点で実験終了とした．

4.3.3 実験結果

各パラメータの全項目に対する実験協力者の回答結果の平均値を図 4 と図 5 に示す．図 4 は，トピックとコミックの関連を推定する際に根拠となった平均単語数を，図 5 は，認知度におけるトピックとコミックの関連に対する平均正解率をそれぞれ表している．

図 4 と図 5 から，条件（2）と条件（3）が条件（1）より参考になった語数が多く，正解率も高いことが確認できる．特に条件（3）は「知らなかったコミック」に対しては 79 %の正解率を示している．さらに Welch の t 検定による検証を行った結果，実験参加者が知らなかったコミックに対する正解率において，条件 (1) と条件 (2) の間 ($t(15) = 0.044$, $p < .05$) 及び条件 (1) と条件 (3) の間 ($t(18) = 0.042$, $p < .05$) には有意な差が存在していたが，条件 (2) と条件 (3) の間 ($t(14) = 0.567$, $p < .05$) では差が存在しなかった．

5. 議論

4. 章で行った実験により，固有表現が混在している特徴語群よりも固有表現の割合が減少している特徴語群がコミックの内容の推測が容易であることが確認された．さらに，固有表現の割合が減少している語群は，被験者がコミックを知らない場合にも既存の語群よりも内容推測が有意であることが確認された．しかし一方では，今回の実験に設けた条件（2）と条件（3）の間には有意な差が確認できなかった．その原因として本稿で作成した固有表現抽出器の性能が挙げられる．固有表現抽出器を用いて抽出を行った結果，条件（2）は条件（1）よりは実質的な固有表現の数において差が存在するが，IDF 値を基準とし摘出を行った条件（3）とは大きな差が存在していなかったと言える．この課題に関しては，今後抽出結果に基づいた再学習や，新たなコーパスを取り入れるなど，抽出器の精度や汎用性を高めていくことで改善を図りたいと考えている．

6. おわりに

本稿では，コミック検索システムに用いる情報には，固有表現というユーザによって情報量が大きく異なるものが含まれている点に着目し，コミックのレビュー文中に存在する固有表現の抽出を試みた．また，Wikipedia やコミックレビュー本文からコーパスを作成し，新たな拡張固有表現タグを追加した上で CRF を用いた機械学習を行い固有表現抽出器を作成した．作成した固有表現抽出器は F 値において 0.526 という性能を持ち，1,000 作品のレビューから 2,000 件以上の固有表現を抽出できたことを確認している．

抽出した固有表現を用いて（1）既存の情報，(2）固有表現を摘出した情報，(3）知名度の低い固有表現のみ摘出した情報，という 3 つの条件を設定し，各条件下で固有表現がコミックの内容推測に与える影響について調査した．その結果，固有表現の割合を減少させることで特徴語群からの理解が容易になることが確認された．

今後は，今回作成した固有表現抽出器から効率的に大量のコーパスを収集し，より精度や汎用性の高い固有表現抽出器の作成を試みる．それに伴い，今回作成した拡張固有表現タグに対してコミックやレビューにおける固有表現の種類に検討する．最後に，実験によって明かになった固有表現の影響を考慮し，コミック検索システムにおける情報の応用方法や情報提示方法の確立について検討する．

参考文献

[1] Blei, D. M., Griffiths, T. L. and Jordan, M. I.: The Nested Chinese Restaurant Process and Bayesian Nonparametric Intference of Topic Hierarchies, *Journal of the ACM*, Vol. 57(2), No. 7 (2010).

[2] Lafferty, J., McCallum, A. and Pereira, F. C.: Conditional random fields : probabilistic models for segmenting and labeling sequence data, *Proceedings of the 18th International Conference on Machine Learning 2001*, pp. 282–289 (2001).

[3] Sekine, S., Sudo, K. and Nobata, C.: Extended Named Entity Hierarchy, *In Proceedings ofthe 3rd International Conference on Language Re-sources and Evaluation*, pp. 1818–1824 (2002).

[4] Yamashita, R., Okamoto, K. and Matsushita, M.: Exploratory Search System Based on Comic Content Information Using a Hierarchical Topic Classification, *Proc. 5th Asian Conference on Information Systems* (2016).