CRF を用いたアニメ関連用語の固有表現抽出

高瀬 真記（東京農工大学　工学部　情報工学科）
古宮 嘉那子（東京農工大学　工学研究院）
小谷 善行（東京農工大学　工学研究院）

Named Entity Recognition for Animation-Related Words Using CRF

Masaki Takase(Department of Computer and Information Sciences
Faculty of Engineering)
Kanako Komiya(Institute of Engineering,　Tokyo University of Agriculture and Technology)
Yoshiyuki Kotani(Institute of Engineering,　Tokyo University of Agriculture and Technology)

１．はじめに

近年，日本のコンテンツ産業は「クールジャパン」という名称のもと注目を集めており，その中でも漫画，アニメーションなどのいわゆるサブカルチャーは，商業的な観点から見ても重要なコンテンツとなりえている．また，アニメーションなどの作品には多くの固有表現が含まれている．それはキャラクターの名前であったり，作中に登場するロボットの名前であったり，作品タイトルそのものであったりである．そして，それら固有表現は商品検索や商品同定，推薦などに利用できると考えられる．しかし，従来の研究ではアニメーション関連用語の固有表現抽出システムは基本的に存在しない．そこで，本研究ではアニメ関連用語に特化した固有表現抽出システムを考える．固有表現抽出手法としては CRF を利用する．

２．関連研究

固有表現抽出は，今まで様々な方法で行われている．その中でも大きく分けるとパターン照合による固有表現抽出と，機械学習による固有表現抽出にわけられる．

パターン照合による固有表現抽出とは，あらかじめ人手で固有表現のパターンを作成し，合致する部分をコーパスから発見することによって行われる固有表現抽出のことである（竹本，福島，山田 (2001)）．パターンとは「さん」や「大学」などの固有表現に付属しやすい文字列を指す．しかし，ルール作成のコストが高く，そこに合致しない固有表現は抽出できないので，助詞を含むタイトルなどが多数存在するアニメ関連用語に用いるのは難しい．

人手でパターンを作成するコストや，更新するコストを解決するために，機械学習による固有表現抽出の研究も行われている．機会学習による固有表現抽出は，学習用のコーパスを用意することで，自動で抽出パターンを学習することができる．機械による手法は SVM(Support Vector Machine)（山田，工藤，松本(2002)）を利用した抽出や文節情報を利用した抽出（中野，平井(2004)）などが存在し成果を上げている．その他にも HMM(Hidden Markov Model)や分類機の逐次適応, CRF(Conditional Random Fields)を利用した固有表現抽出なども一般的である．機械学習の問題点としては人手によるコーパス作成（橋本，乾，村上(2008)）のコストが高いことなどがある．

こうした研究を踏まえ，本稿ではアニメなどサブカルチャーの特殊な固有表現に特化した固有表現抽出について行った．

３．アニメ関連用語

アニメ関連用語の固有表現を抽出する前に，固有表現抽出の対象となるアニメ関連用語を定義する必要がある．本研究ではそれらを「内部の固有表現」と「外部の固有表現」に

分けて定義した．具体的には，それぞれアニメ作品内に登場する固有表現とアニメを制作する製作者などを指す固有表現である．定義した固有表現をそれぞれ表 1 と表 2 に示す．

表 1：内部の固有表現

細かな概念	例
1.1.1 作品内のタイトル	となりのトトロ
1.1.2 作品内のタイトルの略称・通称	トトロ
1.1.3 作品内の人間以外の登場キャラ	ポチ
1.1.4 作品内に登場する必殺技	かめはめ波
1.2.1 作品内の登場人物	宿海仁太
1.2.2 作品内の登場人物の略称・通称	じんたん
1.3.1 作品内の登場神	ゼウス
1.4.1 作品内の登場組織名	現代視覚文化研究会
1.4.2 作品内の登場組織名の略称・通称	現視研
1.4.3 作品内の登場一族	波紋の一族
1.5.1 作品内の登場店	MAHO堂
1.5.2 作品内の登場施設	イゼルローン要塞
1.5.3 作品内の登場サイト	地獄通信
1.6.1 作品内の登場製品	波動エンジン
1.6.2 作品内の登場技術・システム	シビュラシステム
1.6.3 作品内の登場言語	メルニクス語
1.6.4 作品内の固有身分	千翔長
1.6.5 作品内の固有職業	ヴァンパイアハンター
1.7.1 作品内で発生する事件	セカンドインパクト
1.7.2 作品内に登場する計画	人類補完計画
1.7.3 作品内に登場するキャラの特殊行動・任務	オペレーショントルネード
1.8.1 作品内に存在する固有物質	飛行石
1.8.2 作品内に存在する固有生物・種族	アーヴ
1.8.3 作品内に存在する固有の部位	空識覚器官
1.9.1 作品内に存在する固有病	黒鉄病
1.9.2 作品内に登場する特殊能力	幻想殺し
1.9.3 作品内に登場する特殊状態	スーパーサイヤ人

表 2：外部の固有表現

細かな概念	例
2.1.1 作品の製作者	石原立也(監督)
2.1.2 作品の原作関係者	谷川流(原作者)
2.1.3 作品製作関係者	杉田智和(声優)
2.2.1 作品の製作会社	京都アニメーション
2.2.2 作品の放送会社	TOKYO MX
2.3.1 作品の関連商品	まんま肉まん
2.3.2 作品の関連商品に関連する会社	コトブキヤ
2.3.3 メディアミックス関連雑誌等	ジャンプスクウェア
2.3.4 作品の関連サイト	アニメの公式サイトなど
2.4.1 作品に関連したイベント	アニメコンテンツエキスポ
2.4.2 作品に関連したプロジェクト	西尾維新アニメプロジェクト

内部の固有表現は作中に出てくる用語，外部の固有表現は作品の製作者や関連商品販売社などを対象としている．

また，アニメ関連用語として，地名は現実世界に存在する実在の地名とかぶることもあり，アニメを対象とした固有表現としにくいため，対象から除外した．

４．アニメ関連用語の固有表現抽出手法

アニメ関連用語の固有表現抽出は，CRF による系列ラベリングで行う．学習用のコーパスをアニメに関連したコーパスにすることでアニメ関連用語に特化した固有表現をおこなう．固有表現のタグには BIOES 形式を使用した．タグの意味は表 3 に示す.

表 3：タグの意味

タグ	意味
B	その形態素が固有表現の始まりであることを示す
I	その形態素では固有表現が継続していることを示す
E	その形態素で固有表現が終わることを示す
S	その形態素一つで一つの固有表現であることを示す
o	その形態素は固有表現ではないことを示す

利用した素性は「表層」「品詞」「品詞細分類」「文字種」「文字数」の五つである．入力された文章を形態素解析し「表層」「品詞」「品詞細分類」を取り出し，表層から「文字種」と「文字数」を作成する.

５．アニメ関連用語の固有表現抽出実験

提案する手法を用いて，アニメ関連用語の固有表現抽出実験を行った．その際に，形態素解析器として MeCab(http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html)を, 系列ラベリングに CRF++ (http://crfpp.googlecode.com/svn/trunk/doc/index.html)をそれぞれ用いた.

５.１　実験データ

CRF++に学習させる際の学習用アニメコーパスは自身で作成した．対象とした文章は Wikipedia に記事のあるアニメ作品 50 タイトルのあらすじである. その中に含まれる固有表現を先に述べた定義で抜き出し, タグ付けをした. 文字数は 44829 文字で, 表層数は 26948, 固有表現数は 1570 である．その内，S タグで表される固有表現が 742 個で BIE タグで表される固有表現が 828 個である.

５.２　アニメ関連用語の固有表現抽出実験内容

学習用アニメコーパスを五分割交差検定することで評価した．正解のタグとシステムが出力したタグを比較し，S タグの場合は両者が揃った場合，BIE タグの場合は，最初から最後までタグが揃った場合を正解とした.

５.３　アニメ関連用語の固有表現抽出実験結果

表層，品詞，品詞細分類の三つの素性を使った状態の結果をベースラインとし，文字種，文字数，文字種＋文字数の組成を使った状態の結果と合わせて，全部で四種類の素性の組み合わせで出た結果は表 4 のようになった.

表 4：組成ごとの結果

テンプレート	再現率	精度	F値
ベースライン	0.655	0.842	0.737
文字種の素性を追加	0.682	0.844	0.754
文字数の素性を追加	0.667	0.846	0.746
文字種と文字数の素性を追加	**0.687**	**0.848**	**0.759**

全ての値は，文字種＋文字数を使った場合に最大となり，その際，S タグで表される固有表現の精度に関しては，ベースラインから見て 5%有意水準で有意という結果が出た.

６．考察

実験結果から，アニメ関連用語は文字種と文字数の素性を使うとより抽出できていることが分かる．これは，アニメ関連用語には片仮名の単語や，漢字の組み合わせのような単

語が多いことが要因としてあげられる．「スターライトブレイカー」のような形で表される単語は，片仮名であり，さらに表層が区切られにくいため，文字数も多くなりがちである．そういったヒントから，文字種と文字数はアニメ関連用語の固有表現抽出において有用なヒントとなりうると考えられる．しかし，「あの日見た花の名前を僕達はまだ知らない。」や「ジャングルはいつもハレのちグゥ」など，むやみに長いタイトルはうまく抽出できていなかった．この実験では前後二行の素性しか見ていないために，普通の文章と区別がつけられなかったと考えられる．しかし，前後の数を増やすと結果が悪くなる傾向があったので，上手く識別するための素性の改良は今後の課題である．

全体的な結果を見ると，文章が柔らかく，良い結果が出にくい Web 文章を用いたコーパスでの再現率 0.687，精度 0.848，F 値 0.759 という値はよい結果であり，この手法はアニメ関連用語の固有表現抽出に有効である．

７．まとめと今後の展望

本研究では，アニメ関連用語の固有表現抽出を CRF にておこなった．

アニメ関連用語を定義したのちに，自身で学習用アニメコーパスを作成．その学習用アニメコーパスを「表層」「品詞」「品詞細分類」「文字種」「文字数」の素性を持つ形にして，BIOES タグを付け，CRF++に学習させた．アニメ関連用語の固有表現抽出実験を行った結果，BIOES タグによる固有表現の正解率の F 値 0.759 という値を出した．その結果からアニメ関連用語の固有表現抽出にこの手法は有効である．今後，固有表現タグのついた BCCWJ コーパスを利用して，システムの性能をより高めていく予定である．

謝　辞

本研究では，固有表現タグのついた BCCWJ コーパスを参考に素性の設計などを行いました．快くデータをくださった橋本泰一先生に感謝します．

文　献

竹本義美，福島俊一，山田洋志(2001)『辞書およびパターンマッチルールの増強と品質強化に基づく日本語固有表現抽出』情報処理学会論文誌，Vol42，No.6，pp. 1580-1591

山田寛康，工藤拓，松本裕治(2002)『Support Vector Machine を用いた日本語固有表現抽出』情報処理学会論文誌，Vol.43，No.1，pp.44-53

中野桂吾，平井有三(2004)『日本語固有表現抽出における文節情報の利用』情報処理学会論文誌，Vol.45，No.3，pp. 934-941

橋本泰一，乾孝司，村上浩司(2008)『拡張固有表現タグ付きコーパスの構築』情報処理学会研究報告 2008，pp. 113-120

関連 URL

MeCab: Yet Another Part-of-Speech and Morphological Analyzer
http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

CRF++: Yet Another CRF toolkit
http://crfpp.googlecode.com/svn/trunk/doc/index.html