CRFによるWebコーパスからのアニメタイトルに対する訳語の自動抽出

山崎 舞子[1]　森田 一[2]　古宮 嘉那子[1]　小谷 善行[1]
[1]東京農工大学，[2]東京工業大学
50010268059@st.tuat.ac.jp, morita@lr.pi.titech.ac.jp, {kkomiya, kotani}@cc.tuat.ac.jp

1　はじめに

日本のアニメや漫画が海外でも人気となるにつれ，その関連商品も注目を集めている．しかし，海外の人が日本語サイトで売られている商品を検索するのは困難なことである．さらに，このような分野では次々に新しい用語が作り出されるため，静的な辞書で対応するのは難しい．本稿では，音訳値や距離，品詞などを素性として，Conditional Random Fields（CRF）により英語アニメタイトルに対する日本語の訳語候補を抽出する手法を提案する．この際，Webコーパスを使用し，日本語アニメタイトルが英語に翻訳される際，音訳されているものが多く含まれていることに着目した．

2　関連研究

現在，未知語に対する訳語の自動抽出ではパラレルコーパスが必要となるが，使用できる資源が限られている．そこで，Web上のテキストを用いた訳語抽出が注目されている．Webコーパスからの未知語に対する訳語抽出に関しては，これまで多くの研究が行われてきた[3]．

一方，日本のサブカルチャーは世界で注目を集めており，アニメ単語の抽出が商業的に有用である．アニメ用語に着目した固有表現抽出の研究としてCRFによりアニメ関連用語の固有表現抽出を行った高瀬らの研究がある[4]．

本稿では，英語アニメタイトルに対する日本語の訳語候補を自動抽出する．

本研究と関わりの深い研究として，Changらの研究がある[1]．Changらは検索エンジンによって得たスニペットをもとに訳語抽出を行い，音訳，翻訳，距離を素性とし，CRFにより訳語候補を得て，出現頻度によってランク付けを行う手法を提案している．

先行研究がほとんど中国語-英語間で行われているのに対し，本研究では日本語-英語間の対訳対を対象とする．そのため，CRFでタグ付けする際，中国語では文字（漢字）単位でラベル付けを行うのに対し，本研究では形態素解析を行い，形態素ごとにラベル付けを行う．また，アニメタイトルを対象としているため，英語が元にあって和訳した対訳だけでなく，日本語が元にあって英訳した対訳対も多い．日本語が元にあって英訳された訳語は音訳されていることが多いといった特徴があるため，音訳素性に着目した素性設計を行っており，カタカナやローマ字表記も考慮した．

3　システムの概要

システムは以下の手順により英語アニメタイトルに対する日本語の訳語候補の出力を得る．

1. 入力された英語アニメタイトルをクエリとして，検索エンジンによる検索結果上位最大100件のWebページのテキストデータ取得を行う
2. 1で取得した各テキストデータに対して形態素解析を行い，各形態素に対して11種類の素性値を作成する．素性値については後述する
3. あらかじめ学習されたモデルによって各形態素へのラベル付けを行い，訳語候補を抽出する

4　素性の設計

本システムでは，各形態素に対するCRFの素性として以下の11素性を用いた．このうち日本語読み音訳素性値と英語読み音訳素性値については4.1節，4.2節で詳しく説明する．また，音訳距離とは文字列の音としての近さをはかり，小さいほど一致度が高い値とする．

1). 表層
2). 品詞
3). 品詞細分類
4). 文字種（「ひらがな」「カタカナ」「漢字」「アルファベット」「その他」）
5). 英語タイトルと今みている形態素の位置の差（負の数も可）
6). 翻訳辞書にのっているか否か（0 か 1）
翻訳辞書には Wikipedia 日英京都関連文書対訳コーパス [1] を元に GIZA++ [2] を用いて翻訳確率を計算後，0.1 以上の確率がついたペアが格納されている
7). 日本語読み音訳素性値
今みている形態素をローマ字に変換した文字列と英語タイトルの各単語の編集距離を元にした音訳距離．0 から 10 の値
8). 英語読み音訳素性値
今みている形態素をローマ字に変換した文字列と英語タイトルの各単語との動的計画法によるコストを元にした音訳距離．0 から 10 の値
9). 音訳素性値のうち小さい方
10). 括弧の中に存在するか否か（0 か 1）
11). 英語タイトルか否かのタグ

4.1 日本語読み音訳素性値

アニメタイトルには「君と僕（Kimi to Boku)」など，日本語がそのまま音訳されているものが多く存在する．このような訳語を抽出するために，文字列の一致度をはかる日本語読み音訳素性値を導入する．

日本語読み音訳素性値は，英語タイトル中の各単語または連続する二単語と，今見ている形態素，またはその前後と連結した文字列の各組み合わせについて音訳距離を計算し，その最小値を素性値とする．

日本語読み音訳素性値の計算例として，英語タイトル「Dirty pair」として取得した文章「キディグレイド/と/ダーティペア/の/比較」(/形態素の区切りを表す) の中の「ダーティペア」の部分の素性値の計算方法を説明する．

「ダーティペア」はまず,「da-teipea」とローマ字に変換される．その後英語タイトル中の各単語と，隣り合った二つの単語の合成語，すなわち「Dirty」「Pair」「DirtyPair」の三つと「da-teipea」との編集距離がそれぞれ計算される．その中で一番小さい値である「da-teipea」と「DirtyPair」の編集距離 5 を選び，この値が「da-teipea」の文字列の長さである 9 で割られ，0.56 という値が保持される．次に今みている形態素と一つ前の形態素を足しあわせ,「とダーティペア」という形態素を「toda-teipea」とローマ字に変換し，英語タイトルの各形態素との編集距離を計算し，一番小さい「DirtyPair」と「toda-teipea」の編集距離 6 が選ばれる．これを文字列の長さ 11 で割った 0.55 という値を得る．保持しておいた値と比べると今出した値のほうが小さいので，これを 10 倍して四捨五入し,「ダーティペア」の日本語読み音訳素性値は 6 となる．

英語タイトル中の各単語だけでなく，隣り合った二つの単語の合成語を利用することで，この例のように日本語では一形態素だが英語では二単語に分かれるような場合にも対応できる．また，一つ前の形態素と今見ている形態素を足し合わせることで，月/姫（Tsukihime）のように日本語では二形態素に分割できるが，英語では一単語となるような場合にも対応した．

形態素がアルファベットだった場合は，そのままの形態素と英語タイトルに含まれる形態素それぞれとの編集距離をとることで，英語と日本語が混ざった訳語（ビッグ X など）にも対応した．

4.2 英語読み音訳素性値

「フィリックス（firikkusu)」と「Felix」の日本語読み音訳素性値を考えた場合，同じものを指しているのに編集距離が大きくなってしまう．このような英語の音をそのまま日本語にしたような訳語を抽出するために，英語読み音訳素性値を導入する．英語読み音訳値は，今みている形態素と英語タイトルの各単語との音としての近さをはかることができる．

英語読み音訳素性値は，今見ている形態素をローマ字に変換したものと，英語タイトルの各単語を動的計画法を用いて対応づけるが，この手法は動的計画法により音素とカタカナを対応付けた Tsuji らの手法と関連がある [2]．なお，英語読み音訳素性値は注目している形態素の文字種が日本語だった場合のみ計算する．

英語タイトルに含まれる単語のうち一つを英語読み文字列，注目している形態素を日本語読み文字列とする．両者の文字列が一致した場合と英語読み変換規則による文字列の一致がおきた場合にコストは 0 となる．それ以外の場合はコストは 1 に設定した．英語読み変換規則として，英語を日本語に音訳する際の簡単な規則（igh は ai と読む，など）を作成した．

「ドロップ」と「drop」の動的計画法による対応の例を図 1 に示す．英語読み変換規則には日本語文字列

[1] http://alaginrc.nict.go.jp/WikiCorpus/
[2] http://www.statmt.org/moses/giza/GIZA++.html

「do」と英語文字列「d」の対応,「p」と「pu」の対応が記載されているとする．

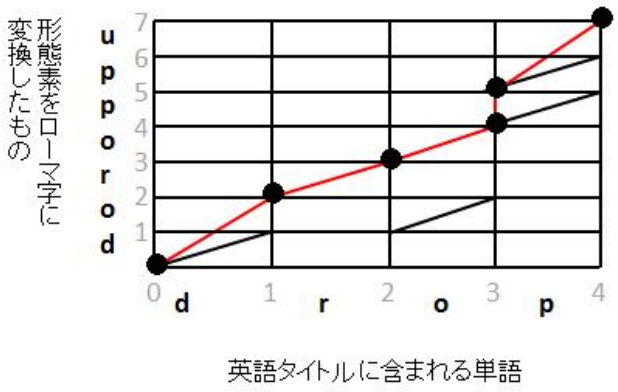

図 1: 動的計画法による対応付け

まず,「ドロップ」は日本語であるため，これをローマ字に直し，「doroppu」とする．英語読み文字列に「drop」を，日本語読み文字列に「doroppu」をセットする．一番左下の 0 の点から対応付けをスタートする．0 の点からは，次の 4 つの点へのパスが与えられる．座標は（英語読み文字列，日本語読み文字列）とする．

・日本語読み文字列「d」の対応がないとする (1,0)
・英語読み文字列「d」の対応がないとする (0,1)
・日本語読み文字列「d」と英語読み文字列「d」の一致による (1,1)
・辞書に記載されている英語読み文字列「d」と日本語読み文字列「do」の対応による (1,2)

0 の点から (0,1) または (1,0) の点に移動するにはコストが 1 かかり，(1,1) または (1,2) に移動するにはコストはかからない．移動する際にはそれぞれの点までのコストも記憶しておく．

次は上で求めた 4 点から移動できるパスをそれぞれ導出する．それぞれの点から移動できるパスを求め，コストを記憶していくことを繰り返し最終的に一番コストの低い経路が日本語読み文字列と英語読み文字列の対応として選ばれる．例の図 1 の場合は最終的に 0 → (1,2) → (2,3) → (3,4) → (3,5) → (4,7) の経路が選ばれ，コストは (3,4) から (3,5) に移動する際にかかる 1 のみとなり，これを以下の式 1 によって計算した値である 1 が「ドロップ」と「drop」の対応を見た場合の英語読み音訳値となる．

$$\text{英語読み音訳値} = \frac{\text{コスト}}{\text{今見ている形態素の文字列の長さ}} \cdot 10 \quad (1)$$

5 評価実験

5.1 実験データ

2012 年の英語版 Wikipedia ダンプデータ [3] より，日英アニメタイトルの抽出を行った．これにより 2134 対のデータを得た．

実験ではこのタイトル対をクエリとして用いて，言語指定を日本語にして検索を行い，各タイトルについて Web ページ最大 100 件を取得した．検索エンジンとしては Google[4] を利用した．

5.2 実験設定

実験を行うにあたって対象とする Web ページを以下のように限定した．

まず，各 Web ページは一つの英語タイトルと，少なくとも一つの正解の訳語を持つこととし，それ以外のページは削除した．今回の実験では，上記の条件を満たす Web ページが一件でも存在し，なおかつ日英で同一のタイトルでない 1282 タイトルを用いた．

一つの Web ページに対して，複数の英語タイトルが存在した場合，一番はじめに出てくる英語タイトル以外については英語タイトルであるというタグを消すことにより対象となる英語タイトルを一つにする操作を行った．

また，英語タイトルと距離が離れすぎている場所には訳語があらわれにくいという考えに基づき，英語タイトルから 25 形態素以内に限定して訳語候補の抽出を行った．

実験において形態素解析器には MeCab[5] を，CRF の実装には CRF++[6] を利用した．

5.3 評価方法

システム全体としては，Wikipedia から抽出した日本語タイトルを正解タイトルとし，正解タイトルとシステムが出力した訳語を比較して，完全に一致するかどうかでシステム全体の評価を行う．一つのタイトルにつき，少なくとも一つの訳語候補を出力した．すべてのタイトルのうち，訳語候補の中に一つでも正解タイトルが含まれているタイトルの割合を示したものをカバレージとし，これを評価した．

[3] http://dumps.wikimedia.org/enwiki/
[4] http://www.google.com/
[5] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[6] http://crfpp.googlecode.com/svn/trunk/doc/index.html

また，参考として，正解タグとシステムが出力したタグを比較して再現率，精度，F 値を求め，評価した．

本実験では，すべての素性を用いたモデルと，すべての素性から「英語読み音訳値」,「日本語読み音訳値」,「音訳値のうち小さい方」の 3 つの音訳に関わる素性を除いたものとの比較を行う．

5.4 実験結果

システムのカバレージを表 1 に，タグの評価結果のマイクロ平均，マクロ平均を以下の表 2 に示す．

表 1: カバレージ

素性	システムがとれたタイトル	カバレージ
全部	922	0.719
-音訳	806	0.629

表 2: マイクロ平均・マクロ平均

	素性	再現率	精度	F 値
マイクロ平均	全部	0.514	0.739	0.606
	-音訳	0.352	0.741	0.477
マクロ平均	全部	0.488	0.588	0.533
	-音訳	0.335	0.531	0.411

表 1 より，すべての素性から音訳を抜いた場合，取得できた訳語の数は 10%近く下がっていることから，従って，音訳素性は訳語抽出において有効であることがわかる．

また，音訳を抜いた場合では，マイクロ平均，マクロ平均ともに精度がほとんど変わらない一方，再現率が悪くなることがわかる．

6 考察

音訳素性の追加によって抽出できるようになったタイトルとして,「（Bismark，星銃士ビスマルク）」,「(Valkyria Chronicles，戦場のヴァルキュリア)」などがあげられる．これは二言語のタイトルに音的なつながりが存在するため，音訳素性が有効に働いたためと考えられる．

逆に，抽出できなくなったタイトルとしては,「(Super big，とんでぶーりん)」や「(Porco Rosso，紅の豚)」などがあげられる．これらは音的なつながりがないため，品詞のつながりや距離などから抽出しなければならないが，学習の結果音訳の重みが大きくなったため，抽出できなくなったと考えられる．

また,「･･･（アニメ）ヤマトよ永遠に　やまとよとわに･･･」というテキストから「BE FOREVER YAMATO」に対する訳語を抽出した際，正解タイトルは「ヤマトよ永遠に」であるが「ヤマトよ永遠にやまと」という誤った出力が得られた．

これは,「YAMATO」に対応する形態素が「ヤマト」と「やまと」の二つ存在し，ともに音訳距離が小さいため，同じ音をあらわす二つの形態素を含む訳語を抽出してしまったと考えられる．このような，英語タイトル中のある部分の読みに対応する日本語の形態素が複数存在する場合を許さないようにすることで，改善できる可能性がある．

また，今回の実験では，抽出できたタイトルが「借りぐらしのアリエッティ」における「アリエッティ」のような一部分では不正解としたが，商品検索などの実用上は十分に目的を果たせる場合があるだろう．

7 おわりに

本稿では，英語アニメタイトルに対する日本語の訳語抽出において，音訳に着目した手法について述べた．実験により，素性として「日本語読み音訳素性値」「英語読み音訳素性値」「音訳素性値のうち小さい方」を入れることで，カバレージが 10%近く向上することを示し，訳語抽出において音訳素性が有効であることを示した．

参考文献

[1] Joseph Z. Chang, Jason S. Chang, and Jyh-Shing Roger Jang. Learning to find translations and transliterations on the web based on conditional random fields. *IJCLCLP*, Vol. 18, No. 1, pp. 19–45, 2013.

[2] Rieko Tsuji, Yoshinori Nemoto, Wimvipa Luangpiensamut, Yuji Abe, Takeshi Kimura, Kanako Komiya, Koji Fujimoto, and Yoshiyuki Kotani. The transliteration from alphabet queries to japanese product names. In *Proceedings of the 26th PACLIC*, pp. 456–462, 2012.

[3] Yuejie Zhang, Yang Wang, and Xiangyang Xue. English-chinese bi-directional oov translation based on web mining and supervised learning. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 129–132, 2009.

[4] 高瀬真記, 古宮嘉那子, 小谷善行. CRF を用いたアニメ関連用語の固有表現抽出. 第三回コーパス日本語学ワークショップ予稿集, pp. 179–182, 2013.