# 別称辞書自動生成ツール：ANDit のオープンソース化

山西 良典[†,a]　　福本 淳一[†,a]

† 立命館大学情報理工学部

a) {*ryama, fukumoto*}*@media.ritsumei.ac.jp*

**概要**　Wikipedia のその表記特徴と構造特徴に基づき，高い適合率で正式名称と別称を対応付けた別称辞書自動生成ツール：ANDit をオープンソースとして公開した．本稿では，ツールの利用方法と共に，ツールによって生成される別称辞書の性能および辞書の拡張性について紹介する．

**キーワード**　別称辞書，オープンソース，検索クエリ

## 1　はじめに

Web 上では，正式名称を短縮した略称や関係性の深い単語で言い換えたりした愛称といった「別称」が使われることが多い．評判分析のための情報検索では，正式名称のみならず別称を用いることで，多くの意見を取りこぼすことなく取得可能になると考えられる．また，レビュー分析などにおいても，このような正式名称と別称を同義のものとして捉えることで，より精度の高いレビュー分析が実現されると考えられる．

本稿では，Wikipedia の構造特徴および表記特徴を利用した別称辞書自動生成ツール（Alternative-Names Dictionary tool:**ANDit**）[1] をオープンソースとして公開した．提案ツールによって生成される別称辞書は一般ユーザが動的に更新する Wikipedia 上の知識を利用することで，動的に更新されると共に一般的に用いられる可能性が高い別称を収録する．これにより，静的な知識体系の中で用いられる正式名称とソーシャルメディア上で実際に用いられることが多い別称の対応付けが実現され，別称を用いたソーシャルメディアから意見情報の自動獲得の効率化支援をねらう．

## 2　ANDit の概要

### 2.1　利用方法

ANDit は以下の手順で利用できる．

1. ツールのダウンロード
   以下のリンクより ANDit をダウンロードする：
   `ttp://www.nlp.is.ritsumei.ac.jp/andit.tml`
2. Wikipedia ダンプデータの準備
   Wikipedia ダンプサイトから以下のファイルをダウンロードする．なお，以下，“xxxxx” はデータがダンプされた日付を示す．ダンプされた日付が同一のものを準備する．
   - jawiki-[xxxxx]-pages-articles.xml
     wiki の記事データ。圧縮され、複数に分割されてアップロードされている．
   - jawiki-[xxxxx]-redirect.sql
     wiki のリダイレクトに関するデータ.
   - jawiki-[xxxxx]-page.sql
     wiki の全ページのタイトル・ページ番号に関するデータ.
3. ANDit の実行
   make_abbreviated_dictionary.pl を perl プログラムとして実行する．ツールを実行する際，第 1，第 2 引数によって作成する辞書の設定を行い，第 3，第 4 引数で使用するダンプデータを指定する．

   **第 1 引数**　別称抽出に用いるキーワードの設定
   「略称」「愛称」「通称」の三種類のキーワードが使用可能であり、各キーワードに対して使用/不使用を設定することができる．

   **第 2 引数**　別称候補の正式名称とのリダイレクト関係の照合通常は，正式名称とその別称候補をリダイレクト関係を照合することで，一般的に用いられている可能性の高い別称のみを抽出する．

   **第 3 引数，第 4 引数**　使用する Wikipedia ダンプデータのディレクトリとタイムスタンプの指定.

### 2.2　抽出処理の概要

ANDit では，Wikipedia のダンプデータを基に以下の手順で別称を抽出する．

1. リダイレクト元とリダイレクト先が対応づいたリストを作成
2. リダイレクト先ページから表記特徴を手掛かりとして別称候補を抽出

表 1　実験に用いた Wikipedia のダンプデータ．左列は 2014 年 5 月 3 日時点，右列は 2015 年 1 月 18 日時点それぞれのデータを示す．

| | 2014 年 5 月 3 日 | 2015 年 1 月 18 日 |
|---|---|---|
| 総ページ数 | 1,824,685 | 1,902,229 |
| リダイレクト元ページ数 | 510,177 | 530,747 |
| リダイレクト先ページ数 | 238,314 | 247,018 |

表 2　適合率評価実験結果：2014 年 5 月 4 日時点のダンプデータを基に生成した別称辞書のそれぞれの適合率評価．数値は%．

| | 略称 | 愛称 | 通称 |
|---|---|---|---|
| 緩い基準での適合率 | 90.5 | 95.3 | 96.0 |
| 厳しい基準での適合率 | 87.3 | 89.6 | 83.0 |

3. 抽出した別称候補とリダイレクト元ページの項目名の完全一致を検証

手順 1 でリストに収録された「ももいろクローバー Z」を例とすると，まず手順 2 で「ももいろクローバー Z」のアブストラクト文からは愛称として「ももクロ」「ももクロちゃん」が抽出される．次の手順 3 では，ももクロ」のみがリダイレクト元の項目名と完全一致するため,「ももいろクローバー Z」の愛称として「ももクロ」のみが辞書に登録される．生成される辞書の例については，文献 [1] を参照のこと．

## 3　ANDit の性能評価

表 1 に示した Wikipedia ダンプデータを用いて別称辞書を生成し，その性能を評価した．

### 3.1　適合率評価

2014 年 5 月 3 日時点での Wikipedia のダンプデータ（表 1 第 2 カラム）に対して，提案ツールを用いて別称辞書の生成実験を行ったところ,「略称」「愛称」「通称」についてそれぞれ 1477 件, 451 件, 359 件が抽出された．ここで，母比率 0.1，標準誤差 0.05 で信頼度 95%を満たすサンプル数を算出し，略称，愛称，通称についてそれぞれ 126，106，100 サンプルを取り出して評価した．20 代の評価者を 3 名用意し，2 段階の基準に従って辞書の性能を評価した．3 名の評価者のうち 2 名以上が適切と判断したものを正例とする緩い基準と，3 名の評価者全てが適切と判断したものを正例とする厳しい基準の 2 種類の評価基準を用意した．

表 2 に，評価実験の結果を示す．同表から，全ての辞書について，緩い基準では 90%以上，厳しい基準でも 80%以上の適合率で別称が抽出されていることが見て取れる．人手を加えずに，自動的に生成された辞書としては十分に高い適合率が示されたと考えられる．別称は固有名詞の増加によって増えたり,「死語」になることで減ったりすることがある．そのため存在する全ての別称の総数を推し量ることは不可能であり，再現率を求めることは難しい．しかしながら，3 種類の手掛かり語によって合計 2,287 件の別称を抽出しており，それらについて比較的高い適合率が確認された．

表 3　拡張性評価実験結果：2014 年 5 月 4 日から 2015 年 1 月 18 日までに追加登録された別称それぞれの適合率評価．数値は%．

| | 略称 | 愛称 | 通称 |
|---|---|---|---|
| 緩い基準での適合率 | 91.6 | 86.5 | 96.0 |
| 厳しい基準での適合率 | 78.5 | 73.1 | 88.0 |

### 3.2　拡張性評価

2015 年 1 月 18 日時点での Wikipedia のダンプデータ（表 1 第 3 カラム）を基に別称辞書を再生成し，2014 年 5 月 3 日時点でのダンプデータを基に生成した辞書との差分を抽出した．その結果，別称辞書には略称は 107, 愛称は 52，通称は 25 データがそれぞれ追加登録されたことが確認された．追加登録された項目全てについて 20 代の評価者を 3 名による評価実験を行った．評価基準には前述の性能評価実験と同様に，緩い基準と厳しい基準を用意した．

表 3 に，評価実験の結果を示す．追加登録された別称についても，全ての別称において緩い基準では 85%以上，厳しい基準では 70%以上の適合率が示された．この結果より，提案ツールは日々増え続ける別称を高い適合率で逐次的に追加可能であることが示された．

## 4　おわりに

本稿では，オープンソースとして公開した別称辞書自動生成ツール：ANDit について，利用方法と抽出手続きを示した．また，ANDit によって生成される辞書について，高い適合率と拡張性を評価実験によって確認した．

Twitter などの文字数制限があるソーシャルメディアでは，別称を用いて意見が記述されることが多い．また，正式名称のみを用いた検索では，bot や広告などが多く取得されてしまい，一般ユーザの意見が埋もれてしまうことがある．提案ツールによってソーシャルメディアからの感情・評判取得の効率化が期待される．

## 謝辞

本稿の執筆にあたって，角野翔大氏の協力を得た．本研究は，一部，人工知能研究振興財団の支援のもと行われた．記して謝意を表す．

### 参考文献

[1] 山西良典，福本淳一，“Wikioedia の特徴表現を利用した別称コーパス生成ツールの開発,” ARG WI2，pp.57–62，2013.