# Webからの名称とその読み情報の自動抽出

森 正樹† 福本 淳一‡

立命館大学情報理工学部メディア情報学科†‡

## 1 はじめに

インターネットの普及により個人が自由に情報を発信することができるようになり，情報発信者以外に読むことが難しい言葉や，ローカルなコミュニティのみで通じる言葉も多く存在するようになってきた．このような読みの難しい言葉に対しては，「よみ」情報が主に丸括弧「( )」を付与することで表されている．

本研究ではWeb上の文書から「よみ」を表す丸括弧を見つけ，括弧の前のどの部分の読みを表しているのかを自動的に判断し，抽出する手法を提案する．特に，英数字の列に対する「よみ」情報について，「よみ」の発音表現の可能な組合わせと対象範囲の照合により，対象範囲を抽出する．これにより，ネット上での新語にも対応可能となる．

## 2 「よみ」情報の分析

基本的に「よみ」情報は丸括弧を用いて付与されているが，どのような場合に丸括弧内が「よみ」と判断できるか，また，「よみ」である場合にもその対象範囲がどのようなものであるかをWeb上の文書を用いて分析する．

### 2.1 丸括弧内の「よみ」情報の分析

丸括弧内の「よみ」に当たる部分が実際に「よみ」となっているか分析を行った．以下にWeb文書の分析から得た「よみ」となっているもの(○)とそうでないもの(×)の例をあげる．

- ○ 近所の函館(はこだて)市場という寿司屋
- ○ このプロジェクトはJAXA(ジャクサ)を
- ○ 車の売り買いなら8710(ハナテン)まで
- × 人体に被害を及ぼす石綿(アスベスト)
- × そんなことはないの(だろう)

分析からは，丸括弧の中の文字種はひらがな，カタカナの場合に「よみ」となっており，「よみ」の対象が漢字などの日本語単語に対してはひらがなが多く用いられ，対象が英数字，記号の場合にはカタカナが多く用いられていた．例中の「アスベスト」のように訳の場合にはカタカナが多く用いられていた．

### 2.2 「よみ」の対象範囲の分析

「よみ」の対象となる範囲として，日本語単語の場合と英数字，記号の場合に分けて分析した．以下に分析を行った「よみ」の例を示す．

黒曜石から作った鏃(やじり)を
雷を発する雷霆（らいてい）を持っています
デザインはKDDIDESIGNING(デザイニング)が
ケータイクレジット「iD(アイディ)」対応

日本語単語の場合は対象となる範囲は主に名詞連続となっており，範囲の特定のために記号などのように明確に範囲を特定できるものと，助詞，動詞，助動詞などの連続する名詞以外のものが範囲の特定に有効であることがわかった．

英字列においては，対象となる英数字列すべてが「よみ」の対象とはならないものがいくつか存在した．上の例では'KDDIDESIGNING'のうち「よみ」の対象は'DESIGNING'のみであり，'KDDI'は対象とはなっていない．このような場合，「よみ」の発音情報から推定することが可能であると考えられる．「デザイニング」という「よみ」のうち「で」から発音として'DE'もしくは'D'が想定され，対象の英数字の2,3,5番目が候補となる．同様にすべての「よみ」の発音から候補を推定することができる．'COME'の'E'のように「よみ」にない文字もいくつか存在するがなるべく多くの読みを含む候補となる範囲を対象として決定することが可能と考えられる．

## 3 「よみ」情報の抽出手法

前節の分析に基づき，Web文書から「よみ」情報を抽出する手法について述べる．本手法は丸括弧内の表現が「よみ」であるかどうかを括弧の前の単語の文字情報によって判断を行う．その際，括弧の中がひらがなの場合，括弧の直前の単語の文字種が漢字，数字，記号のいずれかであれば「よみ」と判断する．また，括弧の中がカタカナの場合，括弧の直前の単語の文字種が英字，数字，記号のいずれかであれば「よみ」と判断する．

次に，抽出範囲の決定手順について述べる．括弧の直前の単語の文字種が漢字，数字，記号のいずれかの場合，丸括弧から左方向に探索し，動詞-自立-，助詞，助動詞，

**Extraction of Reading Information from Web Documents**
Masaki MORI†and Junichi FUKUMOTO‡
Department of Media Technologies, Ritsumeikan University†‡
{m_mori, fukumoto}@nlp.is.ritsumei.ac.jp

句読点, 記号のいずれかが見つかるまでの範囲とする. また, 括弧の直前の単語の文字種が英字, 数字, 記号の場合, 以下の手順により範囲の決定を行う.

(a) かな文字のよみの英字列から可能な組合わせのよみ英字列を生成
(b) 各生成文字列に対して抽出対象文字列と一致する位置を調査し, 一致する位置の数字のうち昇順となるもののみを抽出
(c) 抽出された文字列で最も多くの文字列を抽出範囲とし, 最も範囲の狭いものを抽出範囲と決定

図 1 に「よみ」の発音候補と候補文字列との対応を示す. 「デザイニング」の読みの各文字の「よみ」とそれらを組み合わせた結果を生成し, 候補に対して対象となる 'KDDIDESIGNING' 中の位置を数字で示している. 例えば, 'I' については 'KDDIDESIGNING' の 3,7,10 番目に存在する. 次に, 図 2 に候補中の位置から昇順になるもののみについて範囲の候補を絞り込む. 例えば, 候補として「2 5 6 3 11 7 9 12」の列については昇順になるもののみを選ぶことで「2 5 6 7 9 12」を得ている. このようにして得た列から最も多くの位置情報を含み, 範囲の狭いものとして 4 から 12 の範囲を決定し, 「よみ」の対象として 'DESIGNING' を得ている.

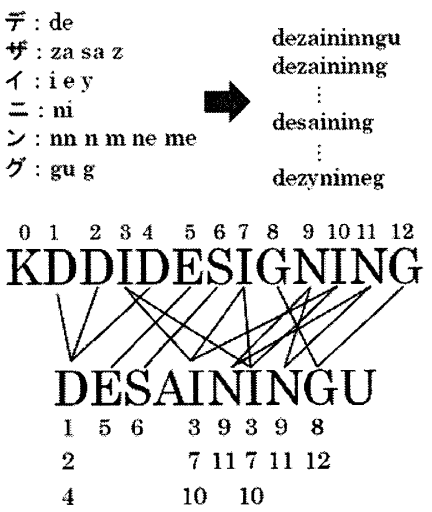


図 1: 「よみ」の発音候補と候補文字列との対応

## 4 実験結果と評価

隅田らの用いた手法 [1] を参考に, Wikipedia [1] 内をクロールして約 1000 件の Web ページから得た「英字 (カタカナ)」のデータ 284 件に対して[2]「よみ」範囲抽出を行った結果. 262 件が正しく認識され, 精度は 92.3% であった.

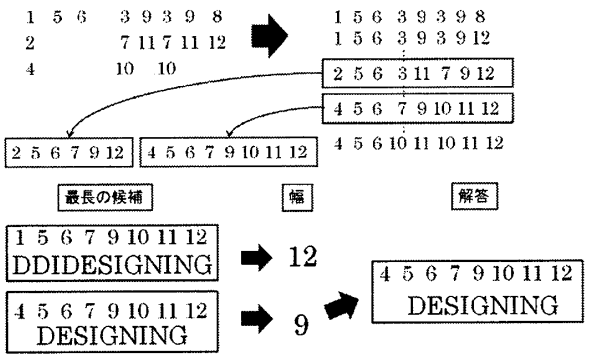


図 2: 候補文字列の対応関係からの範囲の決定

高精度での範囲の決定結果であったことから本手法の有効性を示すことができた. 抽出できなかったものを表 1 に示す.

表 1: 範囲決定の失敗例

| 候補 | よみ | 抽出結果 |
|---|---|---|
| TOYOTANOAH | ノア | NOA |
| Zion | シオン | ion |
| 88Style | ハチハチスタイル | Style |
| AmebaStudio | アメスタ | AmebaSt |
| URL | アドレス | RL |
| XX | ダブルエックス | X |

失敗の原因は, Zion(シオン) について, 想定しない発音として 'z' を「シ」と読むものがあり, これについては発音候補を追加することで対応が可能なものであった. また, 英字の 'h' については発音しないものであり範囲に含まれなかった. 候補の略称となっている「アメスタ」や数字をそのまま読ませるものについては想定外のものであり, 正しく範囲決定できなかった.

## 5 おわりに

本研究では, Web 上の文書から丸括弧の「よみ」がどの部分の読みを表しているのかを自動的に判定する手法を提案した. 今後の課題としては, アルゴリズム改良と数字, 記号のよみへの対応と漢字のよみへの対応がある.

## 参考文献

[1] 隅田, 吉永, 鳥澤, 萬成. Wikipedia からの大規模な上位下位関係の獲得. 第 14 回年次大会発表論文集, 言語処理学会, pp.769-772, 2008.

[1] http://ja.wikipedia.org/wiki/日本の企業グループ一覧
[2] 今回の実験では英数字のみのよみの判定を行った. 漢字のよみについては本範囲設定手法で約 6 割のものに有効であった.