

E-039

# ブログを用いた人物紹介キャッチコピーの自動生成

## A generating method of catchphrases with Blog for a personal introduction

松田 優貴†  
Yuki Matsuda

天沼 博†  
Hiroshi Amanuma

松澤和光†  
Kazumitsu Matsuzawa

### 1 はじめに

近年、技術革新が進みインターネットが生活の一部となった。そのインターネットの普及により、日々の心境や生活などを、ブログという形で誰でも簡単に作成でき、誰にでも紹介できるようになってきている。これは、離れている人や赤の他人との新しいコミュニケーションの形となっており、ブログを通してコミュニケーションする人が増えてきている。しかし、ブログというものがインターネット上で一気に増加したため、数多くのブログが存在するようになった。その数多いブログの中で、より多くの人とブログを通してコミュニケーションするために、より自分のブログに興味を引き付ける必要がある。そこで、本研究では、ブログの自己紹介文を利用し、その人に合ったインパクトのあるキャッチコピーを自動生成するシステムを提案する。

### 2 提案手法

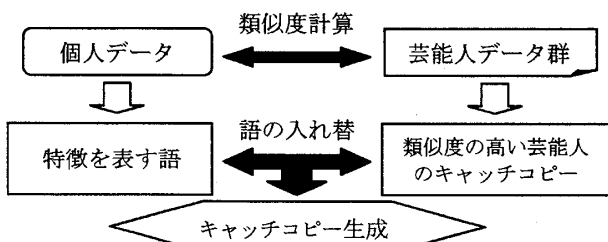
まずキャッチコピーとは、人に強い印象を与えるように物や人の特徴を表現した短い文句。つまり、読み手の目と心を一目でつかむ短い宣伝文のことである。一般的にキャッチコピーは短いものが多く、それは一瞬で読み手を引き付ける必要があるからである。人が一瞬で理解できる文字数は、10文字前後と言われているので、この短い文章の中で、読み手を引きつけなければならない。

そのために本研究では、下記のような昔の芸能人の面白くインパクトのあるキャッチコピーを再利用する。

表1 芸能人のキャッチコピー例

芸能人例	キャッチコピー例
酒井法子	おキャンなレディ
南野陽子	純だね、陽子
山瀬まみ	国民のおもちゃ、新発売

具体的には、ブログの自己紹介文（友人の紹介文を含む）から抽出した個人データと、Web上から集めた芸能人のデータ群から、どの芸能人とその人が似ているか、類似度を計算する。そして、個人データからその人の最も特徴的な語を抜き出し、類似度の高い芸能人のキャッチコピーの語を入れ替え、キャッチコピーを生成する。



### 2.1 人間属性DB

その人の特徴を表す語を抜き出すために、人間の特徴を表す語（性格、外見など）に注目した。色々な人の自己紹介文や友人からの紹介文より、人間の特徴を表す語を抜き出し、22分類 1225語にDB化し、これを人間属性DB①とする。

表2 人間属性DB①の例

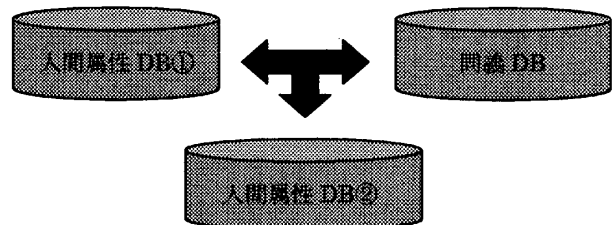
分類例	キーワード例
性格	面白い、明るい、真面目、朗らか・・・
外見	かわいい、かっこいい、イケメン・・・
人物	アキバボーイ、勇者、アネゴ・・・

さらに、人間属性DB①にある語を、分類語彙表を使い意味の似ている言葉同士をまとめ、同義DBとする。

表3 同義DBの例

	分類例
明るい	陽気、朗らか、楽天的、楽しい・・・
かっこいい	イケメン、甘いマスク、モテる・・・
真面目	正直、好青年、ばか正直・・・

そして、人間属性DB①を同義DBで、表2のようにまとめた22分類 580語を人間属性DB②とする。



### 2.2 特徴語の抽出と類似度

#### A. 人間属性DB①を利用（比率で計算）

一般人の自己紹介文（友人からの紹介文を含む）と芸能人のデータからそれぞれ、人間属性DB①にある語を抽出する。このとき、抽出した語をX、その抽出したXの個数をn、そして抽出した語の総数をNとして各語の比率を計算する。

表4 比率の計算例

抽出した語X	抽出した個数n	比率 = n/N
面白い	5個	0.5
優しい	3個	0.3
かっこいい	2個	0.2
抽出した総数N	10個	1

† 神奈川大学大学院工学研究科松澤研究室

この比率の計算を一般人と芸能人の両方に行う。計算した比率を使い、一般人と芸能人一人ひとりの類似度を計算する。類似度は、一般人と芸能人の一致した語の比率をそれぞれ、一般人をS1、芸能人をS2とすると下記で計算する。

$$\text{類似度} = \sum (S1_{各語} \cdot S2_{各語}) \dots (1)$$

**B. 人間属性 DB②と同義 DB を利用 (比率)**

類似度の精度をあげるために、Aで抽出した一般人と芸能人の両方の特徴を表す語を、同義 DB を使い変換を行い、比率を計算しなおす。その後、Aと同様に類似度を計算する。

＜変換例＞

かっこいい 0.22 真面目 0.21 イケメン 0.15 正直 0.1 ……

↓

かっこいい 0.37 真面目 0.31 ……

**C. 人間属性 DB①を利用 (tf\*idfにより計算)**

その人物の抽出した語がよりその人物にとって特有かどうかを見るために、人間属性 DB①にある語を抽出したとき、抽出した語を tf\*idf による重み付けを行う。

- tf : 紹介文に出現するキーワード t の頻度
- idf : キーワード t が全人数に出現する頻度

$$\text{idf} = \log_2 \frac{N}{\text{df}(t)} + 1$$

N : 全人数    df(t) : キーワード t が出現する人数

これで tf\*idf による重み付けを一般人と芸能人の両方に行う。重み付けした tf\*idf の値で、一般人と芸能人一人ひとりをベクトル空間法で類似度を計算する。類似度は、一般人と芸能人の一致した語の tf\*idf の値をそれぞれ、一般人をS1、芸能人をS2とすると下記で計算する。

$$\text{類似度} = \frac{\sum (S1_{各語} \cdot S2_{各語})}{\sqrt{\sum (S1_{各語})^2} \cdot \sqrt{\sum (S2_{各語})^2}} \dots (2)$$

**D. 人間属性 DB②と同義 DB を利用 (tf\*idf で計算)**

Bと同様に、抽出した一般人と芸能人の両方の特徴を表す語を、同義 DB を使い変換を行い、tf\*idfによる重み付けをしなおす。その後、Cと同様に類似度を計算する。

**2.3 キャッチコピー生成**

芸能人のキャッチコピーを再利用する上で、どの言葉の部分にどの分類の語が入るのか表5のようなキャッチコピーDBを構築する。

表5 キャッチコピーDB例

元のキャッチコピー	生成用のキャッチコピー
おキャンなレディ	おキャンな<人物>
純だね、陽子	<性格>ね、<人物>
国民のおもちや、新発売	<性格><人物>、新発売

そして、類似度が大きい芸能人のキャッチコピーを使い、2.2の一般人のデータで比率、もしくは tf\*idf の値が高い分類の語を、キャッチコピーDB の分類にあたる語に変換する。下記に表5を使った例を示す。

例) tf\*idfの値が高い語：真面目(性格)、勇者(人物)

- おキャンなレディ ⇒ おキャンな勇者
- 純だね、陽子 ⇒ 真面目ね、勇者
- 国民のおもちや、新発売 ⇒ 真面目勇者、新発売

**3 評価と結果**

Web上 (mixi) より、一般人 33人分の自己紹介文(友人からの紹介文を含む)をテキスト化し、2.2の手法のうち、今回はA、Bについて、それぞれ芸能人(約90人分)一人ひとりとの類似度を計算し、キャッチコピーを生成した。その中から、手法A、Bの各々で一般人3人分のキャッチコピーをランダムに10個ずつ取り出した計60個を、その人物をよく知る5名に3段階で評価させた。

表6 評価結果

	手法A	手法B
aと評価	20.0%	28.0%
bと評価	32.7%	24.0%
cと評価	47.3%	48.0%
平均2.5点以上	2個	4個

- a : その人に合っていて、おもしろい (3点)
- b : その人に合っているが、普通 (2点)
- c : その人に合っていないし、微妙 (1点)

**4 考察と今後**

表6から、手法AよりBのほうが面白いという結果が得られた。しかし、どちらの手法にしてもcという結果が多い。これは、キャッチコピーを生成したときに言葉の組み合わせがおかしく意味が通らないことが考えられる。また、aという評価が多く得られなかったのは、人によっては頻度の高い語が月並みの語だったため、ありふれたキャッチコピーができてしまったことが挙げられる。これを改善するために、頻度が低くても他の人に出てこない単語を使う(手法C・D)が効果的と考えられる。キャッチコピーDBは係り受けを考慮して自動的に作成するなどの工夫が必要である。また、今回は表記揺れをあまり考慮しなかったので対応が必要である。今後はさらに2.2の手法C・Dによる評価を進めていきたい。また、人に対してだけでなく、商品のキャッチコピー生成などにも応用していきたい。

**参考文献**

- [1] Wikipedia <http://ja.wikipedia.org/>
- [2] 国立国語研究所 編 「分類語彙表」大日本図書 2004