# 頭字語ネーミングの計算モデル

柴田 容子 † 藤井 敦 †† 石川 徹也 ††
† 筑波大学図書館情報専門学群
†† 筑波大学大学院図書館情報メディア研究科
{n166,fujii,ishikawa}@slis.tsukuba.ac.jp

## 1 はじめに

あらゆる物事には名前が必要である。物事に名前を付ける作業を「ネーミング」、名前が必要な物事を「ネーミング対象」と呼ぶ。競合商品との差別化を図るため、名前の独創性が求められている。商標登録やドメイン取得は原則として先願主義であるため、ネーミングの迅速化が求められている。

人手によるネーミングは、「コンセプトの明確化」、「キーワードの収集」、「造語」、「ネーミング案の評価」などの手順に沿って行われる。ただし、各段階において専門的な知識や技能が必要である。そこで、人間の負担を軽減するための支援ツールが存在する。

「名付け親ネーミング辞典[1]」は、ネーミングにおける「キーワード収集」を支援するソフトである。日本語や英語など 10ヶ国語から、ネーミングに適した約 6 万語を収録している。ユーザはイメージを指定してキーワードを検索し、造語の材料として用いることができる。

「ネーミング発想支援システム」[2] は、ネーミングにおいて人間の発想を支援するシステムである。システムには過去の「ネーミング対象の特徴」と「名前」の組が「ネーミング事例」として蓄積されている。そして、ユーザが入力したネーミング対象と似た特徴を持つネーミング事例を検索する。

しかし、既存の支援ツールはネーミングを自動化する度合いが低く、人間の能力に依存する部分が多く残されている。特に、「ネーミング案の評価」を自動化する試みはない。

本研究は、ネーミングの自動化を目的として、計算機によるネーミングのモデル化を行う。さらに、提案したモデルを計算機上のシステムとして実装する。

具体的には、頭字語を作ることで造語を行う。例えば、科学を学ぶ学生に教材やコミュニケーションの場を提供する Web サイト「WISE[2]」の名称は、「Web-based Inquiry Science Environment」というフレーズに対する頭字語である。「WISE」という頭字語によって、このサイトの活動が「賢明である」ことを伝えている。このように、頭字語によるネーミングはフレーズと頭字語の両方に意味を持たせることができる。なお、英語のフレーズと頭字語を対象とする。

[1] http://www.psn.ne.jp/~bds/
[2] http://wise.berkeley.edu/

## 2 頭字語ネーミングの計算モデル

### 2.1 概要

本研究で提案するネーミングの計算モデルは、人間の専門家が行うネーミング手順を模倣している。具体的には、頭字語による造語手法によってネーミング案を生成し、複数の案に対して妥当性を評価し、優先順位を付ける。

ネーミング案の評価では、「フレーズ」に対しては、「文法的に正しさ」と「ネーミング対象を表しているか」という観点に基づいて評価する。「頭字語」に対しては、「印象」、「呼びやすさ」、「覚えやすさ」という観点に基づいて評価し、ネーミング案の妥当性を定量化する。

以上 5 つの観点それぞれを「フレーズモデル」、「コンセプトモデル」、「印象モデル」、「発音モデル」、「単語モデル」を用いて評価する。

本研究で提案する頭字語ネーミングの計算モデルを図 1 に示す。このモデルに基づいてシステムを実装した。

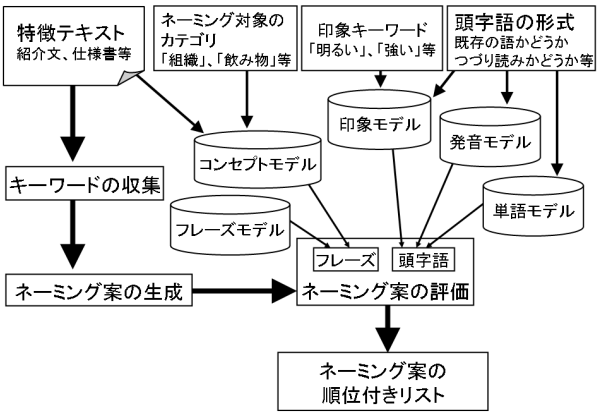


図 1: 頭字語ネーミングの計算モデル

図 1 の各部について説明する。

「特徴テキスト」、「カテゴリ」、「印象キーワード」、「頭字語の形式」の 4 つをシステムへの入力とする。

「特徴テキスト」はネーミング対象の特徴を記述したテキストである。英語以外のテキストは機械翻訳によって事前に英訳する。

「カテゴリ」は、5 種類のネーミング対象カテゴリから指定する。「印象キーワード」として、13 の印象キー

ワードから名前に反映させたい印象を指定する。「頭字語の形式」として、頭字語が既存の語になるべきかどうかを指定する。さらに頭字語を「一文字ずつ読む」か「ローマ字で綴り読みをする」か指定する。

特徴テキストは入力として必須であり、それ以外の入力は任意である。本システムの出力はネーミング案の順位付きリストである。

以下、2.2 節で「キーワードの収集」、2.3 節で「ネーミング案の生成」、2.4 節で「ネーミング案の評価」についてそれぞれ説明する。

## 2.2 キーワードの収集

ユーザが入力した特徴テキストから、造語に用いるためのキーワードを収集する。

「名詞」はネーミング対象そのものを、「形容詞」と「副詞」はネーミング対象の特徴を表す語として用いられることが多い。そこで、名詞、形容詞、副詞をキーワード収集の対象とする。

まず、キーワードとして用いない冠詞、代名詞、数詞、接続詞などの不要語を特徴テキストから削除する。次に、活用形を原形に復元するなどの語尾処理を行ってキーワードを収集する。ここで収集したキーワード群を「一次キーワード群」とする。

次に、一次キーワード群の各キーワードについて同義語を収集し、これらを「二次キーワード群」とする。

品詞付与、不要語の削除、語尾処理、同義語の収集には WordNet[3]を用いた。

キーワード群の各単語に対して、「品詞」と「特徴テキスト内における生起確率」に関するタグを自動的に付与する。品詞はネーミング案の生成で、特徴テキスト内における生起確率はコンセプトモデルで用いる。

二次キーワードの生起確率は、同義語の拡張元となった一次キーワードの多義性を考慮して決定する。具体的には、$N$ 個の意味を持つ単語の各意味は等確率で使用されると仮定する。すなわち、$N$ 個の意味を持つ一次キーワードに対して二次キーワードを追加する場合は、その一次キーワードの生起確率を $\frac{1}{N}$ して、二次キーワードの生起確率とする。

本来は、多義性解消によって、特徴テキスト内で使用されている語義を特定する必要がある。しかし、多義性解消は今後の研究課題とする。

## 2.3 ネーミング案の生成

一次キーワード群と二次キーワード群を用いて、3〜5 語の順列によってフレーズを生成する。

フレーズ生成の際に、各キーワードに付与された品詞タグを参照して、文法的に不適な品詞の並びを含むフレーズを削除する。不適切な品詞の並びとは、「副詞、名詞の並び」、「フレーズ末が形容詞」、「フレーズ末が副詞」の 3 通りである。

次に、生成された各フレーズから頭字語を生成する。頭字語は 3〜5 文字の文字列である。

頭字語の生成において、フレーズ構成語が「exciting」や「extensible」など「ex」で始まる場合は、頭文字の「e」を用いた場合と、先頭から二文字目の「x」を用いた場合についてそれぞれ頭字語を生成する。例えば、「Extensible Markup Language」というフレーズからは、「EML」と「XML」という 2 種類の頭字語が生成される。

また、1 つの頭字語に 2 通りの「読み」を付与する。「読み」はヘボン式ローマ字で表記し、これを「音素綴り」と呼ぶ。2 通りの読みとは、「A」を「エー」、「X」を「エックス」のようにアルファベットを一文字ずつ読む「アルファベット読み」と「PA」を「パ」、「fa」を「ファ」のように読む「綴り読み」である。

頭字語末の文字が綴り読みできない場合は、「d」を「do」、「r」を「ru」のように母音を補完する。母音を補完するパターンは子音ごとに設定しておく。

## 2.4 ネーミング案の評価

ネーミング案の評価では、ネーミング案のフレーズに「フレーズモデル」と「コンセプトモデル」を用いてスコアを付ける。頭字語には「印象モデル」、「発音モデル」、「記憶モデル」を用いてスコアを付ける。ネーミング案を 5 つのモデルで得られたスコアによってソートし、上位から順番に提示する。各モデルのスコアは全て 0〜1 の値をとる。

### 2.4.1 フレーズモデル

フレーズの評価に用いる観点は「文法的な正しさ」である。2.3 節で文法的に不適切なフレーズは削除しているので、ここでは N グラムモデルを用いたフレーズの生起確率によってスコアを付ける。

本研究では単語のトライグラムモデルを用いる。トライグラムモデルは NTCIR テストコレクション[4]英語技術論文抄録 32 万件から、高頻度語 10 万語を用いて学習した。そのため、技術論文に頻出する単語や単語列を含むフレーズに高いスコアが与えられる。

### 2.4.2 コンセプトモデル

「ネーミング対象の特徴を表すかどうか」を評価するために、特徴テキストを用いてフレーズの生起確率を計算する。各フレーズ構成語の特徴テキストにおける生起確率を掛け合わせた値をフレーズの生起確率とする。

式（1）によってフレーズ $W = w_1 w_2 \cdots w_n$ の生起確率を計算する。

$$P(W) = P(w_1) \times P(w_2) \times \cdots \times P(w_n) \quad (1)$$

また、ユーザが任意で指定できる項目として「優先したい語」と「重要度」がある。例えば、「$w_2$ を優先、重要度を 3 倍」と指定した場合は、$w_2$ の出現頻度を 3 倍して $P(w_2)$ を求める。

また、ネーミング対象のカテゴリを指定することができる。そのカテゴリに相応しいフレーズパターンと一致

[3]http://wordnet.princeton.edu/

[4]http://research.nii.ac.jp/ntcir/index-en.html

しているかどうかを検査し、一致した場合は 1、一致しない場合は 0.5 をフレーズの生起確率に掛ける。カテゴリとフレーズパターンの対応を表 1 に示す。

表 1: カテゴリとフレーズパターン

| カテゴリ | フレーズ末の単語 |
|---|---|
| 組織 | 「… association」,「… organization」等 |
| 会議 | 「… conference」,「 council」等 |
| 学会 | 「… institute」,「… society」 |
| 活動 | 「… activity」,「… movement」 |
| 飲み物 | 「… drink」,「… water」 |
| システム | 「… system」 |

#### 2.4.3 印象モデル

印象モデルは、頭字語がネーミング対象の印象に合致しているかどうかを評価する。母音や子音などの音素と印象キーワードの対応表 [1] を用いる。本研究で用いる「印象キーワード」とは、「明るい」、「強い」、「大きい」、「重い」、「軽い」、「鋭い」、「柔らかい」、「元気」、「男性的」、「女性的」、「格調ある」、「知的」、「情感ある」である。表 2 に音素と印象キーワードの対応表を一部示す。

表 2: 音素と印象キーワードの対応表（抜粋）

| 音素 | 明るい | 重い | 軽い |
|---|---|---|---|
| a | 1 | 0.5 | 0 |
| o | 0.5 | 0.75 | 0 |
| e | 0.25 | 0 | 0.75 |
| u | 0 | 0.25 | 0.5 |
| i | 0 | 0 | 1 |
| d | 0 | 0.75 | 0 |
| g | 0 | 1 | 0 |
| k | 0 | 0 | 0 |
| p | 1 | 0 | 0.75 |
| t | 0 | 0 | 0 |

例えば、「GIGA」という文字列をローマ字で綴り読みすると「ギガ」となり、音素綴りは「giga」である。「giga」について「重い」という印象のスコアを求める場合は、音素ごとに「重い」という印象キーワード列のスコアを参照し、全ての音素のスコアを足す。このスコアを音素数 4 で割った値 0.625 を「giga」の「重い」印象スコアとする。5 つの文字列について「明るい」、「重い」、「軽い」という印象に対するスコアを表 3 に示す。ここで、文字列「PICO」を音素綴りに変換すると「piko」となり、「c」が「k」に変わっている点に注意を要する。

表 3: 印象モデルによるスコア付けの例

| 文字列 | 音素綴り | 明るい | 重い | 軽い |
|---|---|---|---|---|
| PICO | piko | 0.375 | 0.1875 | 0.4375 |
| DECA | deka | 0.3125 | 0.3125 | 0.1875 |
| GIGA | giga | 0.25 | 0.625 | 0.25 |
| PETA | peta | 0.5625 | 0.125 | 0.375 |

#### 2.4.4 発音モデル

発音モデルは、頭字語の「呼びやすさ」をモデル化する。筆者らが知る限り「単語の呼びやすさ」の定量化に関する先行研究はない。

本研究では、既存の単語に頻出する音素列は呼びやすい（発音しやすい）という仮説を立てた。そこで、音素の N グラムモデルによって「呼びやすさ」を定量化する。具体的には、ローマ字表記のカタカナ語 110,521 語を用いて音素のトライグラムモデルを学習し、音素列の生起確率を計算する。

#### 2.4.5 単語モデル

単語モデルは、頭字語の「覚えやすさ」を評価する。本研究では、既存の語または既存の語と類似する語は覚えやすいという仮説を立てた。

そこで、WordNet を有意味語（既存の語）の知識ベースとして、頭字語との一致を真 (1) 又は偽 (0) でスコア付けする。ただし、「既存の頭字語」は既存の語として扱わない。既存の語との一致や類似を用いる利点は、頭字語にフレーズの省略形以上の意味を込めることができる点にある。それに対して、既存の頭字語と一致しても、新たな頭字語としての面白みに欠ける。

WordNet から、大文字のみで表記されており、かつ、小文字表記に直したときに同じ綴りの語が存在しない単語を、既存の頭字語とみなした。

#### 2.4.6 スコアの統合

2.4.1〜2.4.5 で説明したスコアは全て 0 以上 1 以下の値をとる。これら 5 つのスコアを掛け合わせて各ネーミング案のスコアを計算する。

## 3 評価実験

### 3.1 概要

ネーミング案の評価に用いるモデルを評価した。

「フレーズモデル」で用いている「文法的正しさ」という観点は、明らかに文法的に不適切なフレーズは排除しているため主観によって評価する余地が少ない。N グラムモデルは機械翻訳や音声認識などで有効性が示されている。

「コンセプトモデル」は体系的な評価をするに至っていない。

「印象モデル」は、音素と印象の対応表（表 2）[1] を用いたネーミング案へのスコア付けを行っており、この表の妥当性は評価が必要である。しかし、予備実験を行ったものの、明確な効果は分かっていない。

以上より、「発音モデル」と「単語モデル」に関する実験結果について 3.2 節と 3.3 節でそれぞれ報告する。さらに、3.4 節でシステムの実行結果について考察する。

### 3.2 発音モデルの評価

実験には被験者を用いて、アルファベットの文字列に付与した「読み」の呼びやすさを判定した。被験者は日本語を母語とする大学院生 2 名である。判定は「呼びやすい」、「やや呼びやすい」、「呼びにくい」の 3 段階である。判定対象の文字列はアルファベット 26 文字をラン

表 4: 発音モデルの評価実験に用いたデータの例

| 文字列 | 読み |
|---|---|
| GZX | ジー ゼット エックス |
| QDAF | キュー ディー エー エフ |
| UNCYP | ユー エヌ シー ワイ ピー |

ダムに 3〜5 文字組み合わせて作成した。実験に使用したデータの一部を表 4 に示す。

被験者が判定したデータについて、発音モデルを適用し、スコアの降順でソートして順位を付けた。各判定値の平均順位を表 5 に示す。

表 5: 発音モデルの評価実験結果

| 判定者 | 判定件数 | 平均順位 | | |
|---|---|---|---|---|
| | | 呼びやすい | やや呼びやすい | 呼びにくい |
| A | 945 | 417.7 | 488.2 | 622.3 |
| B | 1775 | 737.4 | 825.1 | 1017.4 |

実験の結果、音素トライグラムに基づく発音モデルは、「呼びにくい」文字列よりも「呼びやすい」文字列を上位にソートすることが分かった。よって、本研究で提案した発音モデルの有効性が確認された。

## 3.3 単語モデルの評価

単語モデルは、「有意味語は無意味語より記憶に残りやすい」という仮説に基づいている。そこで、有意味語と無意味語の違いが人間の記憶に与える影響について、被験者を用いた認知心理学的手法で実験した。

認知心理学的手法とは、「覚える材料（刺激）を学習し、覚え込む（記銘）。そして、それらを一定の遅延の間覚えておき（保持）、その後で記憶テストを行って刺激を思い出す（想起）」[3] という手法である。

アルファベット大文字表記による 3 文字または 4 文字で構成された「無意味語」と「有意味語」の 2 種類それぞれ 36 語、計 72 語を記憶対象として用いた。実験中に同じ語が複数回出現することは無い。また、「有意味」と「無意味」とは、英語としての意味が有るか無いかである。被験者は日本語を母語とする大学生 5 名である。

被験者は 3 語ずつ提示される文字列を記憶し、一秒おいた後、記憶した文字列を書き取る。被験者 5 名から得られた書き取りの平均正解率を表 6 に示す。

表 6: 単語モデルの評価実験結果

| | 有意味語 | 無意味語 |
|---|---|---|
| 平均正解率 | 0.29 | 0.63 |

実験の結果、有意味語の記憶率は無意味語の約 2.2 倍となり、有意味語は無意味語より記憶に残りやすいことがわかった。すなわち、単語モデルの有効性が確認された。

## 3.4 実行例と考察

特徴テキストとして、本研究「頭字語ネーミングの計算機によるモデル化」[4] の抄録を用いた。すなわち、本システムに対するネーミング案を作成した。

優先する語は「naming（ネーミング）」と「acronym（頭字語）」とし、「重要度を 3 倍」と指定した。カテゴリは「システム」と指定した。印象キーワード、頭字語の形式をいくつか変えて実行し、結果について考察した。

### 3.4.1 良いネーミング案の例

- **Naming Analysis Model Easy System : NAMES**

頭字語の形式を「既存の語」と「アルファベット読み」、印象キーワードを「鋭い」と指定し、頭字語の文字数 5 文字の 13 位に出現した。頭字語が「名前」という意味の単語と一致している。フレーズには「naming」や「model」などの語が含まれている。

- **Acronym Naming System : ANS**

頭字語の形式を「既存の語」と「綴り読み」、印象キーワードを「軽い」と指定し、頭字語の文字数 3 文字の 2 位に出現した。頭字語が「答え（answer）」の略記と一致している。フレーズには「acronym」や「naming」などの語が含まれている。

以上のネーミング案は、頭字語とフレーズの両方が、本システムの特徴を良く表している。

### 3.4.2 良くないネーミング案の例

- **Assignment Model Acronym Support System : AMASS**

頭字語の形式を「既存の語」と「アルファベット読み」、印象キーワードを「軽い」と指定し、頭字語の文字数 5 文字の 9 位に出現した。このフレーズには、ネーミング対象の特徴を表さない「assignment（指名）」が含まれている。

理由は 2 つ存在する。一つ目は、特徴テキスト中の単語に対して、多義性解消を行っていないためである。「assignment」は「naming」の同義語として追加された。しかし、「naming」には「ネーミング」や「指名」などの意味があり、「assignment」は本システムの特徴を表さない「指名」に関する同義語である。

二つ目の理由は、「assignment model」の語順が技術論文中に頻出するため、フレーズモデルのスコアが高くなったことである。そのため、同様の語順が他のネーミング案にも頻出した。今後はネーミング対象のジャンルに関するコーパスを用いてフレーズモデル（単語のトライグラムモデル）を適応させる必要がある。トライグラムモデルの適応は音声認識で有効性が示されている。

# 4 おわりに

今後の研究課題は、システムの総合的な評価、キーワードの多義性解消、フレーズモデルの適応、頭字語以外の造語によるネーミングへの応用である。

# 参考文献


[1] 岩永嘉弘. ネーミングの成功法則. PHP, 2002.

[2] 今野宏. ネーミング発想支援システム. 特開平 5-282357（特許出願), 1992.

[3] 高野陽太郎（編）. 認知心理学 2　記憶. 東京大学出版会, 1995.

[4] 柴田容子. 頭字語ネーミングの計算機によるモデル化. 卒業論文, 筑波大学, 2006.