B-011

# クラス名の単語列に対する品詞列ごとのクラス名数の定量的調査
# A Quantitative Study of the Number of Class Names Having Each POS Sequence in Word Sequences of Class Names

福田宏樹 † 早瀬康裕 † 北川博之 †
Hiroki Fukuda Yasuhiro Hayase Hiroyuki Kitagawa

## 1. はじめに

ソフトウェア開発者は識別子の名前を用いてその識別子の指すものが何であるかを表現する．そこで，識別子名から得た知識を開発や保守の補助に利用する研究が多く提案されている [5, 7, 10]．

本研究では，多数の OSS プロジェクトから得た Java のクラス名の構造を，品詞の並びという側面から定量的に調査する．具体的には，クラス名を構成する単語列に品詞判定を自動で行って得られる品詞列について，品詞列ごとに，その品詞列を持つクラス名の数を数える．品詞判定では品詞列を一つだけ出すのではなく，可能性のある品詞列を全て出すようにするため，クラス名の数は 2 通りの数え方で数える．一つは品詞列が一つに定まるクラス名の数であり，もう一つは品詞列が候補の一つになっているクラス名の数である．本研究では，インタフェース名はクラス名の一種と見なす．インタフェース名はクラス名とほぼ同じ構造を持っていると考えたためである．

一般の開発者らが識別子にどのような構造の名前を付けているか調べることには 2 つの意義がある．まず，識別子名の構造に関する知識は，識別子名を用いて開発や保守を補助する手法を新たに考案する手がかりとなったり，既存の手法の精度向上に寄与すると期待される．次に，この知識は，それ単体でも，一般の開発者にとって命名の参考となる有用な情報である．

## 2. クラス名の品詞判定方法

クラス名から，そのクラス名が持つ可能性のある品詞列の集合（候補品詞列集合）を自動で得る方法を，図 1 に沿って説明する．この方法は，大きく分けて，クラス名を単語分割して単語列を得る処理，単語に候補品詞判定を行って候補品詞集合を得る処理，候補品詞集合同士の直積集合をとって候補品詞列集合を得る処理の 3 段階から成る．単語分割と候補品詞判定については，それぞれ 2.1 節と 2.2 節で詳しく説明する．

### 2.1. 単語分割

クラス名から単語列を得るために単語分割を行う．クラス名が `UPPER_SNAKE_CASE` か `lower_snake_case` の場合はアンダースコアで分割し，`UpperCamelCase` か `lowerCamelCase` の場合は小文字と大文字の境界で分割する．なお，この 4 種の何れでもないクラス名は，単語分割の方法が自明でないため調査対象から除外する．また，末尾の単語が Test,Tests,Impl の何れかであるクラス名は，末尾の単語を外したものが本来の名前と考えられるため，調査対象から除外する．

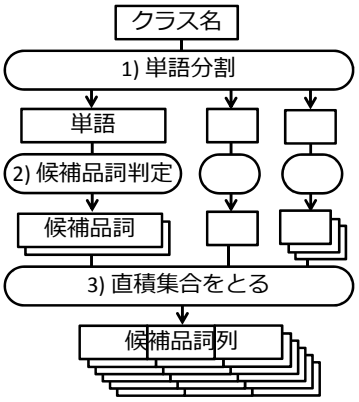


図 1: クラス名の候補品詞列集合を得る方法

### 2.2. 候補品詞判定

単語から候補品詞の集合を得るために候補品詞判定を行うアルゴリズムを，Algorithm 1 に沿って説明する．出力は，名詞単数形，名詞複数形，動詞原形，動詞三単現，動詞現在分詞形，動詞過去分詞形，形容詞原級，副詞，前置詞の 9 品詞のうち 0 種類以上である．

前置詞以外の 8 品詞の判定には，WordNet[8] の Java 向けライブラリである JWNL[1] の機能を利用する (2–8 行)．WordNet は英単語の原形と，名詞・動詞・形容詞・副詞の 4 品詞との，多対多の関係を記録している．JWNL は 4 品詞それぞれに対応した原形化器を持っている．そこで，JWNL を利用して単語の原形が WordNet 上にその品詞として存在するか調べることで，まず，変化形情報を考えない品詞の判定を行う (3–4 行)．次に，単語とその原形を比較して，どの変化形なのかも調べる (5 行)．

前置詞の判定は，Wikipedia の前置詞リスト [2] から a, an, save といった通常前置詞として用いられない語を除いたリストを作成しておき，そのリストに含まれる語かどうか調べるという方法で行う (9–11 行)．

## 3. クラス名の品詞判定結果に対する調査

SourceForge.net[3] から，19,768 の OSS プロジェクトのトップレベルクラスとトップレベルインタフェース計 3,875,281 個を集めた．このうち，3,419,891 個のクラス名に対して品詞判定ができたため，それらのクラス名に対して調査を行った．

まずクラス名全体の傾向を見るために，クラス名の単語数の分布を図 2 に示す．全クラス名のうち，4 単語以下が 92.6%を占めており，大半のクラス名がこの範囲に収まることが分かる．

クラス名の構造を調査することを目的として，品詞判定により品詞列が一意に定まったクラス名について，品詞列を数えた結果を表 1 に示す．Java の命名規約ではク

†筑波大学, University of Tsukuba

**Algorithm 1** 候補品詞判定

**Input:** 単語
**Output:** 候補品詞集合
候補品詞集合 := ∅
**for all** $p \in$ { 名詞, 動詞, 形容詞, 副詞 } **do**
  原形 := JWNL の $p$ 用原形化器 (単語)
  **if** 原形が WordNet に $p$ として存在 **then**
    $p'$ := 単語と原形を比較して変化形を算出
    候補品詞集合 += $\{p'\}$
  **end if**
**end for**
**if** 単語 $\in$ 前置詞リスト **then**
  候補品詞集合 += { 前置詞 }
**end if**

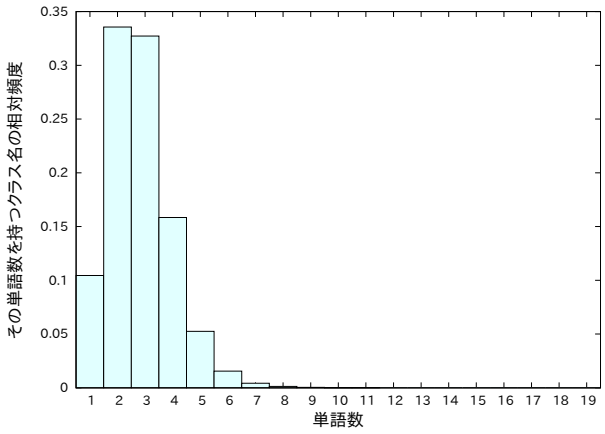


図 2: クラス名の単語数の分布

ラス名を名詞節とすることが推奨されており [4]，実際に名詞節となるクラス名が多いことが確認できる．「動詞原形 名詞単数形」のように，名詞節でないと考えられるクラス名も少数ながら存在する．この品詞列に対応する具体的なクラス名には，`ValidateHelper`，`CreateUser` などがあった．前者は本来 `ValidationHelper` と名詞を使うべきところを動詞で代用したもの，後者は user を create するという動作を表現するために命名されたものと推測される．

調査対象の全クラス名について，可能性のある品詞列を全て数え上げた結果を，表 2 に示す．この数え方においても，順位に多少の変化はあるが，上位の傾向は表 1 同様であることが分かる．クラス名に使われる単語のうち 34.7%は名詞単数形と動詞原形が同形の単語だった．表 1 よりもその他が多い理由は，動詞原形を含む品詞列が多く数えられているためであると考えられる．

## 4. 関連研究

Butler らは Java のクラス名の構造を調査した [9] が，本研究と異なり，単語数ごとの構造には注目しておらず，単語の変化形も区別していない．Høst らは Java のメソッド名とメソッドの実装との関係を調査し [6]，その知識をメソッドの改名候補と改名案の提示へ応用した [7]．Caprile らは C の関数名の構造を調査し，その知識を関数名の自動改名リファクタリングへ応用した [5]．

表 1: 品詞列が一つに定まるクラス名の数

| 品詞列 | クラス名数 |
|---|---|
| 名詞単数形 名詞単数形 | 221,175 |
| 名詞単数形 | 135,821 |
| 名詞単数形 名詞単数形 名詞単数形 | 70,652 |
| 名詞複数形 | 26,341 |
| 名詞単数形 名詞複数形 | 26,046 |
| 動詞原形 名詞単数形 | 14,323 |
| 形容詞原級 名詞単数形 | 13,168 |
| 名詞単数形 名詞単数形 名詞単数形 名詞単数形 | 10,759 |
| 名詞複数形 名詞単数形 | 10,110 |
| 形容詞原級 | 9,403 |
| その他 | 85,176 |

表 2: 品詞列が候補の一つになっているクラス名の数

| 品詞列 | クラス名数 |
|---|---|
| 名詞単数形 名詞単数形 | 738,984 |
| 名詞単数形 名詞単数形 名詞単数形 | 538,608 |
| 動詞原形 名詞単数形 | 355,323 |
| 名詞単数形 動詞原形 名詞単数形 | 291,707 |
| 名詞単数形 動詞原形 | 288,591 |
| 動詞原形 名詞単数形 名詞単数形 | 277,379 |
| 名詞単数形 | 249,190 |
| 名詞単数形 名詞単数形 動詞原形 | 207,238 |
| 名詞単数形 名詞単数形 名詞単数形 名詞単数形 | 198,719 |
| 動詞原形 動詞原形 名詞単数形 | 151,714 |
| その他 | 6,338,343 |

## 5. まとめ

本研究では，Java のクラス名を構成する単語列について，品詞列ごとのクラス名数を調査し，単語数が少ないクラス名が多いこと，名詞節を持つクラス名が多いが少数の例外もあることを確認した．今後は，この知識を活用してクラス名の推薦を行う予定である．

**謝辞** 本研究は科研費 40423090 の助成を受けた．

## 参考文献

[1] JWNL. http://sourceforge.net/projects/jwordnet/
[2] List of English prepositions - Wikipedia. http://en.wikipedia.org/w/index.php?title=List_of_English_prepositions&oldid=581398803
[3] SourceForge.net. http://sourceforge.net/
[4] Sun Microsystems. Code Conventions for the Java Programming Language. http://www.oracle.com/technetwork/java/javase/documentation/codeconvtoc-136057.html
[5] B. Caprile and P. Tonella. Restructuring program identifier names. ICSM 2000.
[6] E. W. Høst and B. M. Østvold. Software Language Engineering, chapter The Java programmer's phrase book. Springer-Verlag, 2009.
[7] E. W. Høst and B. M. Østvold. Debugging Method Names. ECOOP 2009.
[8] G. A. Miller. WordNet: A lexical database for English. Comm. ACM 38(11), 1995.
[9] S. Butler et al. Mining Java class naming conventions. ICSM 2011.
[10] Y. Kashiwabara et al. Recommending Verbs for Rename Method using Association Rule Mining. CSMR-WCRE 2014.