

小規模OSSにおけるUnknownの コスト削減手法のための調査

松浦 遊也

背景

- ソフトウェア開発において、コスト削減のために、OSSは頻繁に再利用されている
- OSSを再利用するにはソースファイルに記述されているライセンスを遵守する必要がある
- ソースファイルは膨大に存在し、理解するのに時間がかかるので、コスト削減のためにライセンス判定ツールが存在する

ライセンス判定ツール(Ninka)



ライセンスルール

ライセンスを特定するための正規表現

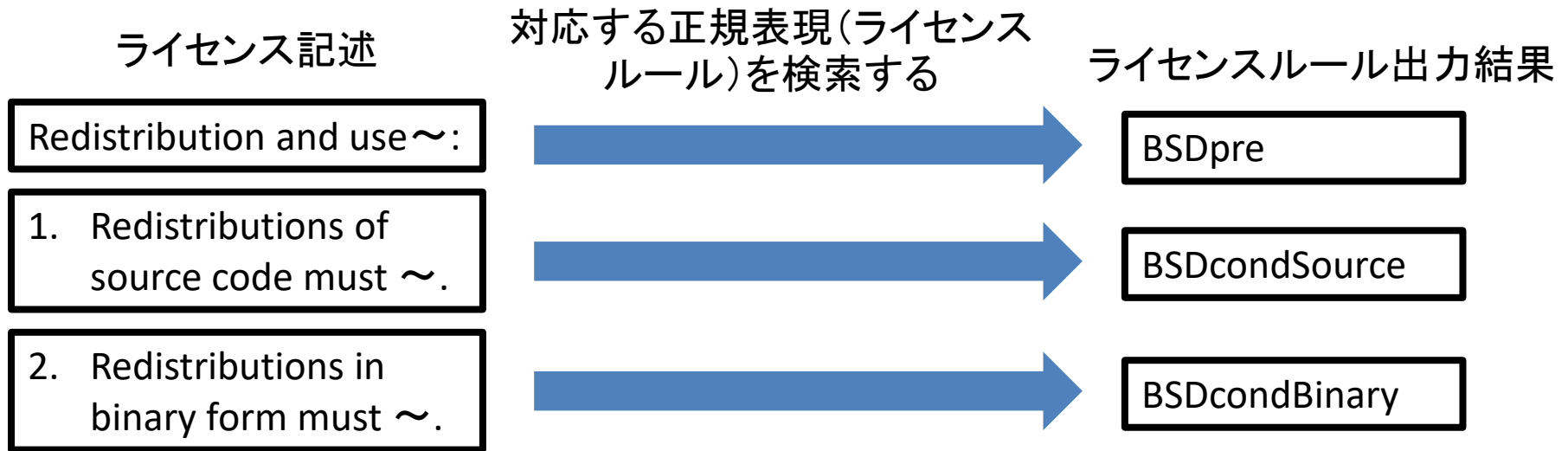
実際のBSDのライセンス記述

Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

ライセンスルール: BSDcondSource

Redistributions? of source code must retain the (above)?copyright notice, this list of conditions(,)? and the following disclaimer(, without modification)?:

Ninkaのライセンス判定:例



読み込んだ文の構成

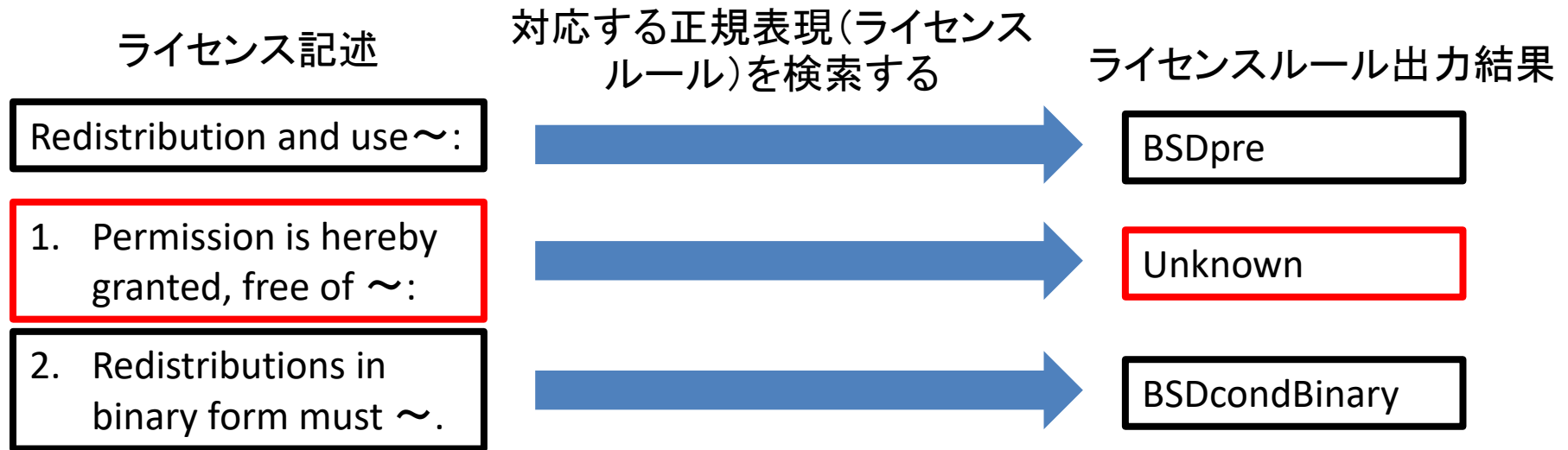
BSDpre, BSDcondSource, BSDcondBinary

BSD2の構成

BSDpre, BSDcondSource, BSDcondBinary

“BSD2”

Ninkaのライセンス判定:例



読み込んだ文の構成

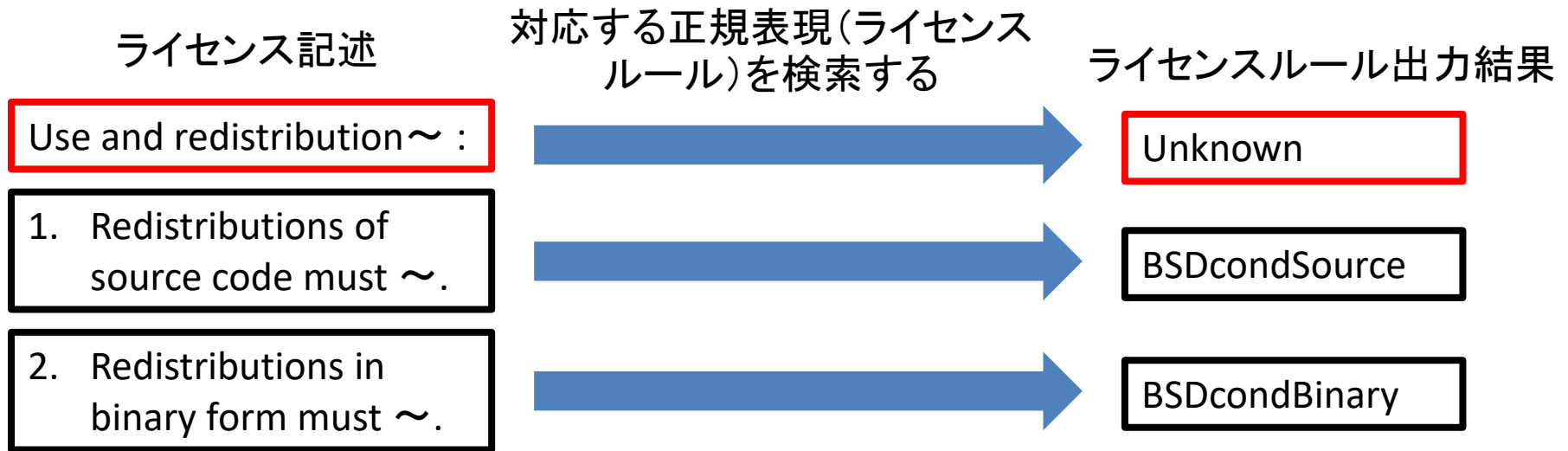
BSDpre, **Unknown**, BSDcondBinary

BSD2の構成

BSDpre, BSDcondSource, BSDcondBinary

“Unknown”

Ninkaのライセンス判定:例



読み込んだ文の構成

Unknown,BSDcondSource,BSDcondBinary

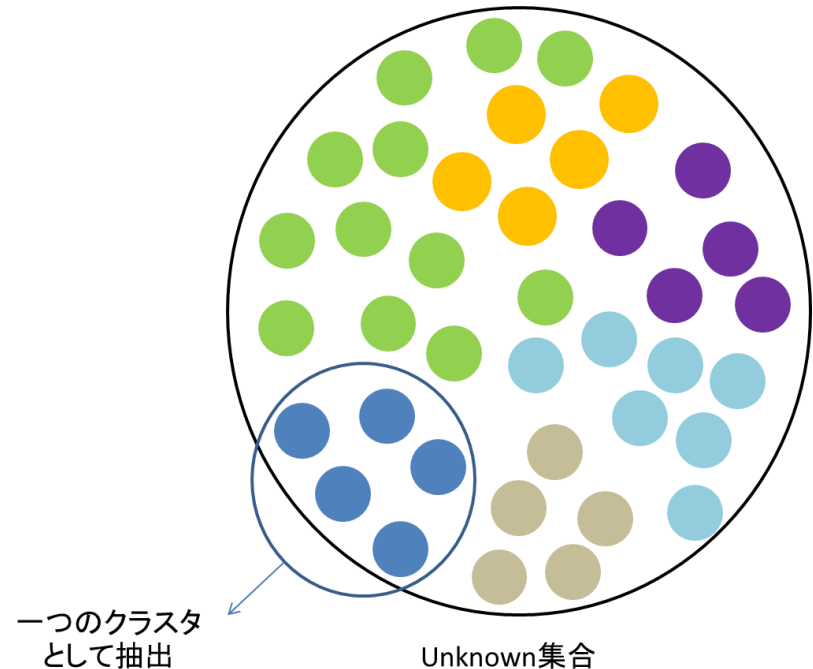
BSD2の構成

BSDpre,BSDcondSource,BSDcondBinary

“Unknown”

関連研究

- Unknownと判定されたライセンスから、単一ライセンスとして抽出するためのクラスタリング手法[1]



ライセンスルールを作成し
Ninkaにフィードバックする研究

[1]Yunosuke Higashi and Yuki Manabe and Masao Ohira **Clustering OSS License Statements toward Automatic Generation of License Rules**, Proceedings - 7th International Workshop on Empirical Software Engineering in Practice, IWESEP 2016,pp30-35

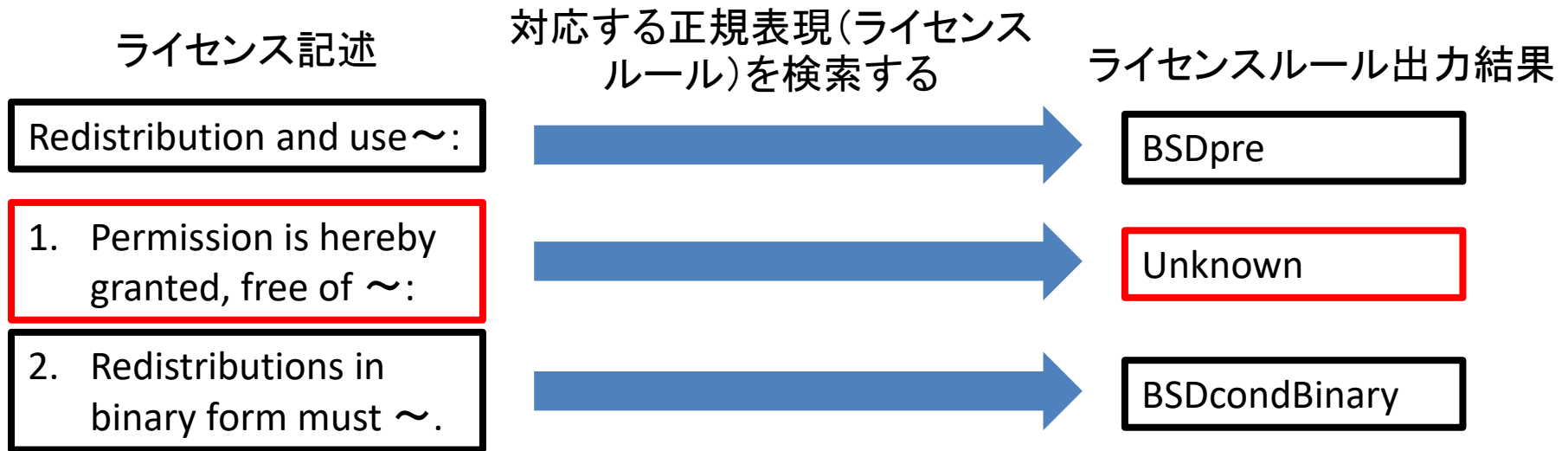
関連研究の支援対象

- 関連研究の支援対象
 - 大規模OSS (Debianなど)
 - Unknownが非常に多いので、ライセンスルールを作成できる
- 関連研究の支援対象外
 - 小規模OSS
 - Unknownがそれほど多くないので、ライセンスルールの作成が困難

目的

- 小規模OSSにおけるUnknownの目視するコストを下げる

Ninkaのライセンス判定:例



読み込んだ文の構成

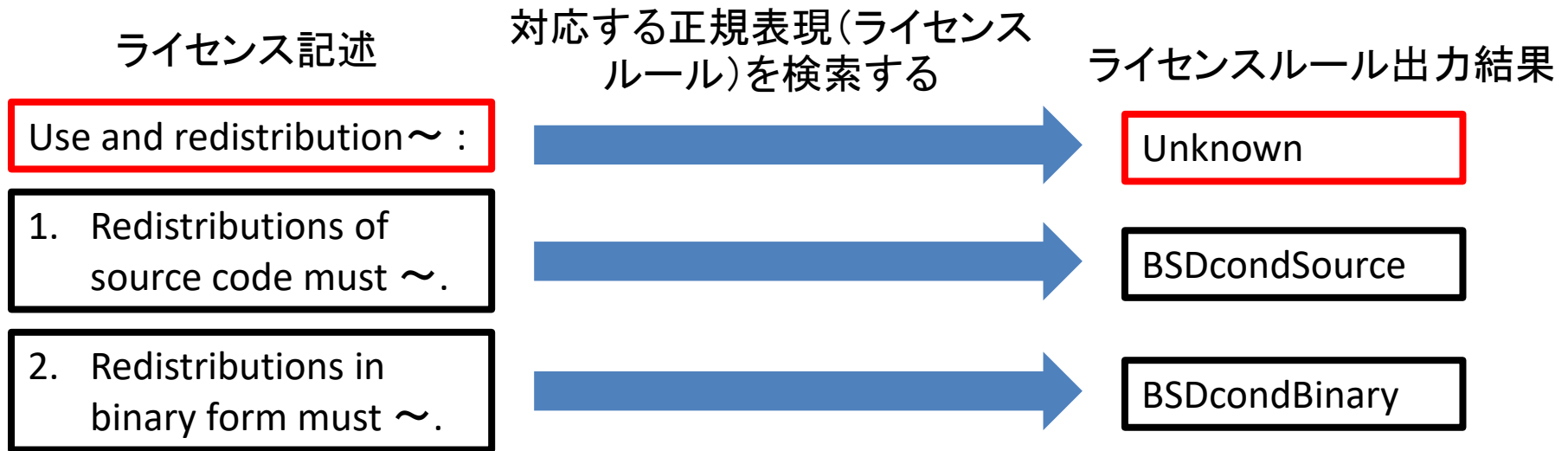
BSDpre, **Unknown**, BSDcondBinary

BSD2の構成

BSDpre, BSDcondSource, BSDcondBinary

“Unknown”

Ninkaのライセンス判定:例



読み込んだ文の構成

Unknown, BSDcondSource, BSDcondBinary

BSD2の構成

BSDpre, BSDcondSource, BSDcondBinary

“Unknown”

提案手法

1. ライセンス記述からUnknownと判定されたライセンス文を抽出する
2. 抽出した文から意味的に似ているライセンス文を除外する
3. 結果を出力する

問題

- Unknownと判定されたライセンス文の総数、既知のライセンス文と意味的に似ているライセンス文がどの程度存在しているのか不明

調査内容

- Debianの12725の各ソフトウェアパッケージからランダムに1つのファイルを抽出しNinkaにかけUnknownと判定されたものを用いる
(2838)
- Unknownと判定されたライセンス記述を文単位で分割(9487)

調査内容

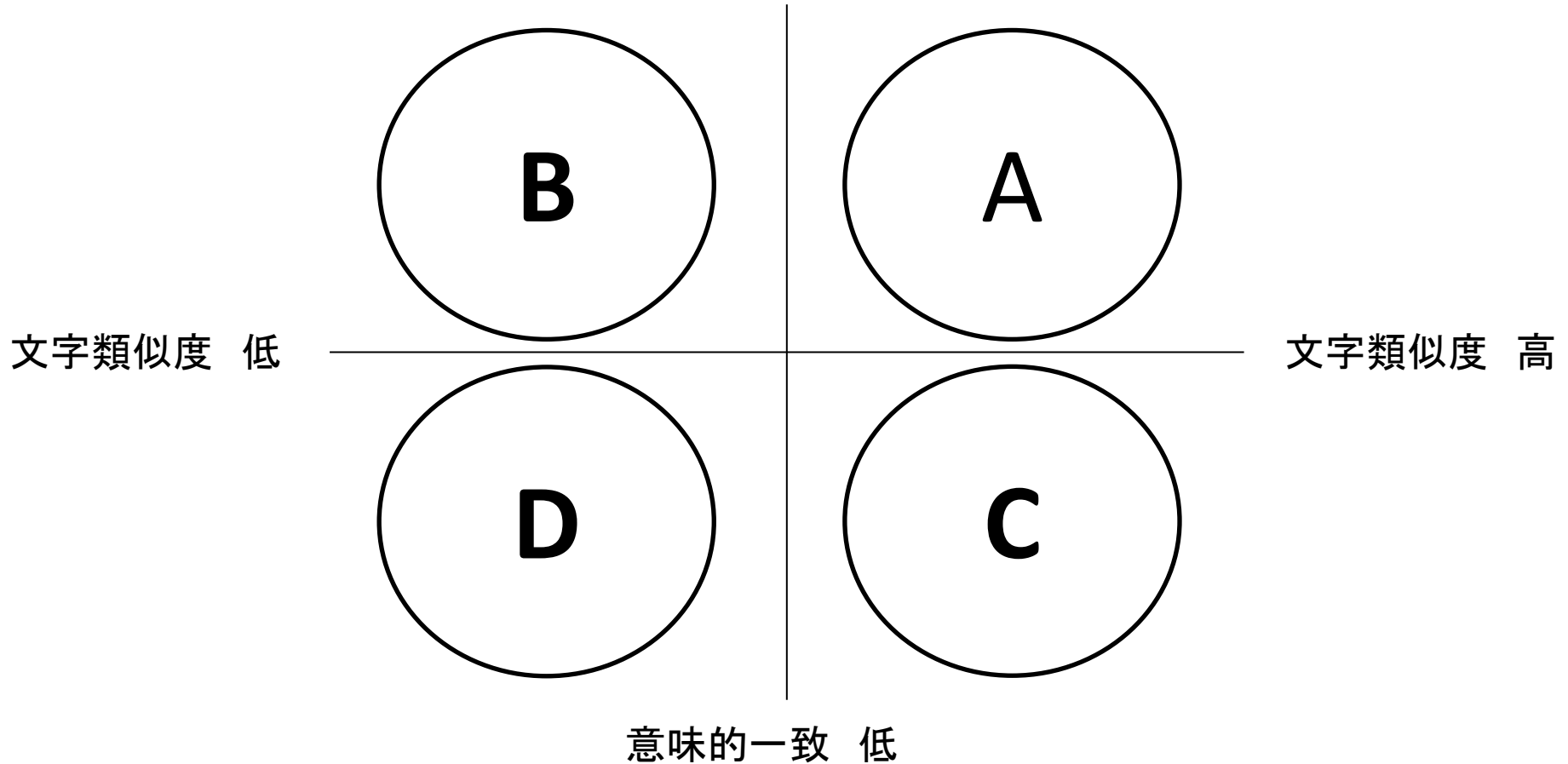
- 行毎に分割したデータセットからUnknownと判定されたライセンス文を抽出(5164)
- Unknownと判定された文の重複(1618)を削除(3546)

評価方法

- Unknownと判定された文3546とNinkaのライセンスルール551を目視で比較していき、意味的な違い、文字列的な違いで分類する

Unknownのタイプ

意味的一致 高



タイプ

- タイプA
 - 文章、意味が既知ライセンスと類似している
- タイプB
 - 文章は類似していないが、意味が既知ライセンスと類似している
- タイプC
 - 文章は類似しているが、意味が既知ライセンスと類似していない
- タイプD
 - 文章、意味が既知ライセンスと類似していない

Aタイプのライセンス文：例

- 実際のライセンス文(Debian)
- This distribution is free software; you **may** redistribute it and/or modify it under the same terms as Perl itself.
- Ninkaのライセンスルール
- `^([^,;]+)` is free software; you **can** redistribute it and/or modify it under the same terms as Perl itself\$

Bタイプのライセンス文：例

- 実際のライセンス文(Debian)
- Redistribution of this file is permitted under the GNU Public License.
- Ninkaのライセンスルール
- ^([^,;]+) can be freely redistributed under the terms of the GPL\$

Cタイプのライセンス文：例

- 実際のライセンス文(Debian)
- This module is free software; you can redistribute it **or** modify it under the same terms as Perl itself.
- Ninkaのライセンスルール
- ^([^,;]+) is free software; you can redistribute it **and/or** modify it under the same terms as Perl itself\$

Dタイプのライセンス文：例

- 実際のライセンス文(Debian)
- This software is subject to the provisions of the Zope Public License, Version 2.1 (ZPL).
- Ninkaのライセンスルール
- ^This Source Code Form is subject to the terms of the Mozilla Public License, v<dot>2<dot>\$0: