

自然言語処理によるソフトウェア設計調査の改善

キャノンイメージングシステムズ株式会社

井越 宣芳

igoshi.nobuyoshi@mail.canon

開発における問題点

実際のソフトウェア開発業務では、新規プロジェクトを立ち上げる業務ばかりではなく、既存プロジェクトをバージョンアップする業務も多い。プロジェクトに機能追加するために、**様々な種類の大量の仕様書や設計書やソースコードから、機能に関連するものを探す**ことが必要。調査漏れはバグを生み品質・生産性が低下する。

手法・ツールの適用による解決

自然言語処理を用いて、「追加機能」の文章と関連する文章（類似文章）を仕様書や設計書やソースコードから網羅的に検索できるようにする。

1. 各種ファイルから文章を抽出するPCアプリ作成。
2. 自然言語処理モデルを用いて、日本語／英語の区別なく、文章を解析（ベクトル化）する。
3. 類似判定結果を一覧表示するWebアプリ作成。

類似文章検索フロー

文章抽出 → 文章解析 → 類似判定

一連の処理をUI操作するツールを作成

文章抽出
PCアプリを作成

文章解析・類似判定
Webアプリを作成



ファインチューニングし類似判定精度向上

文章解析では、XLM-RoBERTaベースの多言語対応した自然言語処理モデルでベクトル化した。ファインチューニングし、日本語/英語区別なくピンポイントで、ソフトウェアで用いる文章の類似判定精度を向上させた。

文章 1	文章 2	向上前	向上後	差
ディレクトリを作成する。	Create a directory.	0.63908	0.96323	0.32416
僕が君と一緒に住んだことは君の人生観に影響を与えたと思う。	I think my living with you has influenced your way of living.	0.93677	0.93580	-0.00097
公園のベンチに乗った白い大きな鳥を見ている人物が二人います。	道を傘をさした女性が歩いています。	0.25656	0.26752	0.01096
A flock of birds hovered over the sandy beach.	Birds are flying at a beach.	0.86382	0.86601	0.00219

類似度の最大値は 1.00

類似文書検索ツールの動作確認

- ・類似判定対象の文章数： 101,472 （文章抽出PCアプリで取得）
- ・検索した追加機能文章： 「登録したAAAAAやBBBBBをPCにダウンロードする」

順位	類似度	文章	分類
1	0.870488	AAAAAをアップロードする	ソースコード
3	0.780081	XX可能なAAAAAインストールセットリストを検索する	仕様書
5	0.766320	AAAAAのBBBBBのインポート	ソースコード
8	0.744148	AAAAAインストールセットのID	設計書
16	0.735895	AAAAA(英語) Download Set file path	ソースコード

比較的高い類似度で日本語/英語区別なく、関連する仕様や実装を検索できた。

⇒ 機能追加の参考にしやすい。

この例では特に、仕様書にない関連実装を英語コメントから見つけることができた。

成果

様々な種類のファイルを横串で日本語/英語区別なく類似文書を検索できた。

一度、抽出文章をベクトル化しておけば、Webアプリで、何度も短時間で類似判定が可能。

本修了制作のツールを展開し、既存の取り組みに加えることで、生産性向上が期待できる。