重複する料理レシピの書き換えのタイプに着目したコーパスの構築と分析
Developing Near-duplicate Recipe Corpus and its Analysis Based on Permutation Types

学籍番号：201621617
氏名：島田　理紗子
Risako SHIMADA

料理レシピサイトにおいて，重複する料理レシピの存在は，ユーザの利用を阻害する要因となりうる．ユーザがレシピを検索した際に，検索結果に似た内容で調理手順が同じレシピが提示されると，探したいレシピの選択に時間がかかってしまう．この問題を解決するためには，重複する料理レシピの判別を行う必要がある．本研究では，重複する料理レシピ（重複レシピ）を判別するためのコーパスを構築し，重複レシピにおける書き方の傾向について分析する．

これまで，重複レシピの判別手法として，調理手順の類似度が用いられてきたが，類似度だけでは，判別が難しいレシピが存在している．本研究では，重複レシピは，元のレシピを書き換えて作られるものとして，その書き換えのタイプに着目し，料理，材料，調理手順の共通性によってタイプを区別した，完全重複，部分重複 A，部分重複 B，部分重複 C，部分重複 D，非重複 A，非重複 B の 7 つのタイプを定義する．そのうち，部分重複 B は，「同一料理で，材料の変更に伴う調理手順の書き換えが行われているが，工夫がみられないレシピ」と定義する．

コーパスは，n-gram 類似度 ($n = 3$) を用いて抽出した重複レシピ候補を，4 名でアノテーションして，3 名以上が同じタイプを付与した 87 件のレシピで構築した．なお，判定者間一致度を計算したところ，Fleiss's kappa は，0.737 （かなりの一致）であった．構築したコーパスのうち，部分重複のいずれかが付与されたレシピを用いて，調理手順の書き換え箇所のレシピ用語の種類について，分析を行った．その結果，部分重複 B では，“食材” の書き換えがほぼすべてのレシピで行われており，約 70% のレシピで “調理者の動作” の書き換えが行われていた．また，この結果に基づき，部分重複 B のレシピについて，“食材” に対応する用語を除いた調理手順を用いて，n-gram 類似度（$n = 3$）の計算を行った．その結果，調理手順の類似度の平均が 0.600 から 0.704 に上昇し，書き換えられるレシピ用語を削除することにより，重複レシピの判別精度を向上させる可能性を示した．

本研究の貢献を以下にまとめる．

(1) 重複レシピについて，7 つのタイプを定義し，コーパスを構築した．
(2) 部分重複のタイプごとに，書き換えられるレシピ用語の種類に違いがあり，それらを考慮して類似度を計算することで，重複レシピの判別精度を向上させる可能性を示した．

研究指導教員：関　洋平
副研究指導教員：芳鐘　冬樹