BERT を利用した文書間類似度と単語埋め込み間の対応に着目した重複レシピの検出

Method for Detecting Near-duplicate Recipes Focused on Document Similarity using BERT and Correspondence between Word Embeddings

学籍番号：201821613
氏名：小邦　将輝
Masaki OGUNI

投稿型レシピサイトには，調理手順テキストなどのレシピの構成要素が他のレシピと重複したレシピ，すなわち「重複レシピ」が存在する．重複レシピの存在により，レシピ検索に余分な時間を費やす等，レシピサイトのユーザビリティへの影響が懸念される．そのため，重複レシピを検出し，それらを削除する等の対処が必要である．剽窃に関する問題は，学術論文や小説，楽曲など，幅広い分野が抱えている．これらの背景を踏まえると，本研究で行う重複検出技術の提案には，様々な応用範囲がある．

これまで，重複レシピの判別手法として，レシピ間の文字 3-gram 集合の Jaccard 係数を用いた手法が提案されてきた．しかし，レシピには特有の言い回しや省略された表現が多く存在する．また，ユーザ投稿型レシピサイト上のレシピは，投稿者によって料理用語の表記が異なり，誤字や脱字が含まれている．そのため，文字 3-gram の類似性を手がかりとした場合，検出できない重複レシピが存在する．

本研究では，文章の意味的な一致を判定するタスクなどにおいて成果を挙げている事前学習言語表現モデル BERT を用いて抽出した調理手順テキストの埋め込み表現間の距離と，単語間の対応に基づき文書間の距離を算出する Word Mover's Distance（WMD）を組み合わせることで重複レシピを検出する手法を提案する．これにより，調理手順テキストの意味についても捉えることができ，単語間の対応についても考慮した重複レシピの検出が行える．提案手法では，BERT を用いて調理手順テキストから埋め込み表現を抽出し，調理手順テキストの埋め込み表現間の距離を基に重複レシピペア候補をランキングする．続いて，調理手順テキストが一致しているが材料が相違しているレシピペアの重複レシピとしての誤検出防止を目的として，材料相違数によるレシピペアのフィルタリングを行う．最後に調理手順テキストの埋め込み表現間の距離上位の重複レシピペア候補について WMD を算出し，調理手順テキストの埋め込み表現間の距離と WMD を基に重複レシピペア候補をリランキングする．

実験では，提案手法と複数の比較手法による重複レシピの検出を行い，各手法の重複レシピの検出精度を比較した．重複レシピの検出数を基に評価を行った結果，材料相違数のフィルタリングによる効果が定量的に示された．また，WMD によるリランキングを行う手法と行わない手法の間で有意水準 1% の有意差を確認した．

本研究の貢献は以下の 3 点である．

（1）BERT により抽出した調理手順テキストの埋め込み表現に基づく重複レシピ検出手法の提案
（2）材料相違数によるレシピペアのフィルタリング手法の提案
（3）WMD を用いた重複レシピペア候補のリランキング手法の提案

研究指導教員：関　洋平
副研究指導教員：高久　雅生