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

筑波大学
図書館情報メディア研究科
2020年3月
小邦　将輝

目 次

表 目 次

図 目 次

第1章　はじめに

1.1　本研究の目的と意義

本研究では，投稿型レシピサイト上に存在する重複したレシピを検出するための手法の提案およびその有効性の検証を行う．レシピサイトにおける重複したレシピの存在による，レシピサイトのユーザビリティへの影響が懸念されている．例えば，ユーザがレシピサイトを利用してレシピを検索した際，検索結果に調理手順テキストが重複しているレシピが複数表示され，かつその中にユーザが求めているレシピがない場合，ユーザが目的のレシピを選択する過程において，余分な時間的コストを掛けることになる．また，こうした事象が繰り返し発生し，ユーザがレシピサイトを利用しなくなった場合，レシピサイトの運営者にとっても不利益を被る結果に繋がる．そこで本研究では，重複したレシピ，すなわち重複レシピの検出手法を提案し，上記の問題を解決する上での一助とすることを目的とする．

重複レシピの検出に関する研究は，著者らの先行研究 [21, 33, 20] に加えて，久保ら [39] などによって行われてきた．著者らの先行研究 [21] および久保らの研究では重複レシピ検出の際に，2 つのレシピの文字 3-gram 集合間の Jaccard 係数を基に，重複レシピの検出を行った．しかし，重複レシピの投稿者の中には，調理手順テキストの言い換えや書き換えを行った上でレシピを投稿する者もいるため，上記の手法のように，文字列的な特徴のみを考慮した単純な手法では検出できない重複レシピが存在する．そこで，単語や文書を高次元の実数ベクトルで表現する埋め込み表現を用いて，意味的な特徴についても考慮した手法を提案する．埋め込み表現では，単語や文書を出現文脈を考慮して学習することにより，類似した単語や文書には類似したベクトルが割り当てられる．本研究では，事前学習言語表現モデルである Bidirectional Encoder Representations from Transformers（BERT）[4] および単語間の対応に着目して文書間の距離を算出する手法である Word Mover's Distance（WMD）[13] を用いて，調理手順テキストの埋め込み表現間の類似性だけでなく，単語の埋め込み表現間の対応にも着目して重複レシピの検出を行う手法を提案する．また，材料相違数によるフィルタリングを取り入れ，調理手順

テキストが一致している場合でも，材料が相違しているレシピペアを重複レシピとして検出することを防止する．

剽窃に関する問題は，投稿型レシピサイトのみならず，学術論文や小説，楽曲など，幅広い分野が抱えている．また，近年ではソーシャル・ネットワーキング・サービス（SNS）上でも，他人の投稿を剽窃した投稿が問題視されるケースもある．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といえる．

1.2 本研究の概要

1.2.1 レシピサイト

料理を作成する際，以前は料理本や料理雑誌といった書籍を用いてレシピを探す機会が多かった．対して，昨今ではインターネット環境の普及に伴い，レシピサイトを利用してレシピを探す機会が増加した．クックパッド[1]による調査[2]の結果，料理をする際に最も参考にする情報源として，レシピサイトが挙げられている．この結果は，今やレシピサイトが料理を探す上で書籍に取って代わる存在になっていることを示唆している．

レシピサイトには，クックパッドや楽天レシピ[3]などの投稿型レシピサイトや，みんなのきょうの料理[4]といった料理番組の Web サイト，キッコーマンの「ホームクッキング[5]」や味の素の「レシピ大百科[6]」など，食品メーカーが提供する情報サイトなどがある．これらのレシピサイトの中で，ユーザが投稿したレシピを Web サイト上に掲載する「投稿型レシピサイト」にはユーザによって投稿されたレシピが掲載されている．

日本を代表する投稿型レシピサイトの一つである楽天レシピには，175 万件を超えるレシピが投稿されている（2019 年 12 月現在）．投稿型レシピサイトの特徴として，幅広いジャンルのレシピが掲載されていること，同じ料理に対しても複数の調理方法が掲載されていることが挙げられる．例えば，楽天レシピにおいて「肉じゃが」をクエリとして検索を行った場合 4,733 件,「かぼちゃの煮物」をクエリとして検索を行った場合 2,787 件のレシピがヒットする．このように検索結果に様々なレシピが表示されることから，ユーザは自らの料理スキルや，好みの味付けに応じてレシピを選択することができる．

[1] クックパッド: https://cookpad.com
[2] https://cf.cpcdn.com/info/assets/wpcontent/uploads/20140306000000/pr130723-survey.pdf（閲覧日: 2019 年 10 月 20 日)
[3] 楽天レシピ: https://recipe.rakuten.co.jp/
[4] https://www.kyounoryouri.jp/
[5] ホームクッキング (キッコーマン): http://www.kikkoman.co.jp/homecook/
[6] レシピ大百科 (味の素): https://park.ajinomoto.co.jp/

こうした特徴により，料理スキル等にかかわらず，様々なユーザに投稿型レシピサイトは利用されているものと考えられる．

投稿型レシピサイトは上記のような利便性を誇る一方で，調理手順テキストや料理画像といったレシピの構成要素が他のレシピと完全に一致，もしくは大部分が一致するレシピが存在する．本研究では，このようなレシピのことを重複レシピ (1.2.2 項参照) と呼ぶ．

1.2.2 重複レシピ

本項では，重複レシピがどのようなレシピであるかを述べる．また，重複レシピが投稿される要因についても言及する．

「**重複レシピ**」とは，1.2.1 項で述べた通り，調理手順テキストや料理画像といったレシピの構成要素が他のレシピと完全に一致，もしくは大部分が一致するレシピのことを指す．調理手順テキストには，レシピの作成者の個性が表れるため，レシピ作成者ごとに異なる記述がなされる．そのため，使用材料が同一もしくは大部分が一致しており，同一の料理について記載したレシピであっても，レシピの作成者が異なる場合，読み手からみると全く異なるレシピに映る．よって，レシピの構成要素が他のレシピと完全に一致，もしくは大部分が一致する重複レシピが偶然の一致により作成されるとは考え難い．

なお，本研究では重複レシピの剽窃元となったと考えられるレシピを「**オリジナルレシピ**」と呼ぶ．すなわち，重複レシピはオリジナルレシピを模倣して作成されたとみなす．また，重複レシピとオリジナルレシピのペアを「**重複レシピペア**」と呼ぶ．

重複レシピペアの例を表 1.1，1.2 に示す． 表 1.1 に示した重複レシピペアでは，タイトル，使用材料，調理手順テキストのすべてが完全に一致している．投稿日に着目すると，上側のレシピが 2018 年 9 月 4 日，下側のレシピが 2018 年 7 月 20 日となっている．すなわち，上側のレシピが下側のレシピを剽窃して作成されたと考えられるため，上側のレシピが重複レシピ，下側のレシピがオリジナルレシピとなる．

重複レシピには表 1.1 のような，レシピの構成要素のすべてが完全に一致しているものばかりでなく，表 1.2 のように，調理手順テキストの一部で言い換えや書き換えが行われているものも存在する．本研究における重複レシピの判定基準に関しては，4.2 節において詳細を述べる．

投稿型レシピサイトは多様なユーザが利用するため，検索結果の多様性 [31] が要求される．しかし，重複レシピが多数存在することにより，検索結果に調理手順テキストが重複しているレシピが複数表示される等，検索結果の多様性に影響を及ぼす恐れがある．

表 1.1: 重複レシピペアの例 (完全一致)

重複レシピ	タイトル	玉ネギ卵ツナ炒め
	使用材料	・玉ねぎ・ツナ・マヨネーズ・塩こしょう ・卵
	調理手順テキスト	(1) 玉ねぎを切る (2) フライパンにマヨネーズを入れ玉ねぎがしんなりするまで炒める溶き卵を加え、混ぜる (3) 卵が固まったらツナ塩コショウを入れ混ぜて完成
	レシピ投稿日	2018 年 9 月 4 日
オリジナルレシピ	タイトル	玉ネギ卵ツナ炒め
	使用材料	・玉ねぎ・ツナ・マヨネーズ・塩こしょう ・卵
	調理手順テキスト	(1) 玉ねぎを切る (2) フライパンにマヨネーズを入れ玉ねぎがしんなりするまで炒める溶き卵を加え、混ぜる (3) 卵が固まったらツナ塩コショウを入れ混ぜて完成
	レシピ投稿日	2018 年 7 月 20 日

そのため，重複レシピを検出し，レシピを排除することや，質の低いレシピについて，レシピ検索を行った際のページランキングの順位を下げるといった対策を講じる必要がある．楽天レシピの事業部に対する聞き取り調査[7]を行った結果，重複レシピの存在はレシピサイトの運営者も認知しており，その存在を問題視していることが明らかになった．1 日に 500 件が投稿される楽天レシピにおいて重複レシピが占める割合は高くない．しかし，重複レシピの数は着実に増加しており，サービスの品質に大きな影響を与える前に，対策が必要になる．

重複レシピを検出するためには，新たに投稿されたレシピに対して，それまでに投稿されているレシピの中から重複したレシピが存在するのかを調査する必要がある．すなわち，新たに投稿されたレシピを「重複レシピ候補」とすると，重複レシピ候補に対応する「オリジナルレシピ」を発見しなければ，新たに投稿されたレシピが「重複レシピ」であるかどうか判別できない．また，同調査において，1 日に 500 件程度のレシピが楽

[7]実施日: 2018 年 9 月 19 日

表 1.2: 重複レシピペアの例 (一部書き換え)

重複レシピ	タイトル	ベーコンチーズかぼちゃコロッケ
	使用材料	・かぼちゃ・たまねぎ・ベーコン・小麦粉 ・粉チーズ・塩こしょう・卵・パン粉・油
	調理手順テキスト	(1) かぼちゃをレンジで 5 分くらい加熱し、マッシュする (2) フライパンで細切りにしたベーコン、みじん切りにしたたまねぎを炒める (3) 1 と 2、塩こしょう、粉チーズを混ぜ合わせ、小判形に丸める (4) 小麦粉、卵、パン粉の順に衣をつけ、180 度の油できつね色になるまで揚げる
	レシピ投稿日	2018 年 9 月 13 日
オリジナルレシピ	タイトル	かぼちゃベーコンチーズコロッケ
	使用材料	・かぼちゃ・ベーコン・チーズ・たまねぎ ・塩こしょう・小麦粉・卵・パン粉・油
	調理手順テキスト	(1) かぼちゃをざく切りにし、レンジで 2〜3 分くらい加熱し、マッシュする (2) フライパンでみじん切りにしたベーコン、たまねぎを炒める (3) 1 と 2、チーズ、塩こしょうを混ぜ合わせ、小判形にする (4) 小麦粉、溶き卵、パン粉の順に衣をつけ、油できつね色になるまで揚げる
	レシピ投稿日	2017 年 4 月 28 日

天レシピに投稿されていることが明かされた．1 日に 500 件のレシピが，新たに投稿される現在の状況では，人手でこの作業を行うことはきわめて困難である．そこで本研究では，重複レシピの疑いがあるレシピを自動検出することで，人手で重複レシピ検出を行う上での一助とする．

重複レシピが投稿される要因として，投稿型レシピサイトがレシピ投稿者に対して付与する報酬が挙げられる．楽天レシピを例にとると，投稿したレシピが掲載された場合，1 レシピにつき，50 円分の報酬が付与される[8]．多くの報酬を得るためには，多くのレシピを投稿する必要があるため，単に多額の報酬を得ることを目的に重複レシピを投稿し

[8]楽天スーパーポイント付与について (楽天レシピ): https://recipe.rakuten.co.jp/rules/point.html

ているレシピ投稿者が存在すると考えられる．

著者の先行研究 [35] による調査では，1 時間に 30 件以上のレシピを投稿したユーザが存在することが明らかになった．楽天レシピのシステムでは，1 レシピを作成するためには，少なくとも 5 分程度は時間を要すると考えられることから，1 時間に 30 件以上ものレシピを投稿することはきわめて不自然である．このことは，多額の報酬を得ることを目的として，意図的に多数の重複レシピを投稿するユーザの存在を裏付けている．著者らの先行研究 [21, 35] では，短時間に多数のレシピを投稿したユーザを重複レシピの投稿が疑われるユーザとして重複レシピの検出の手がかりとした．しかし，重複レシピの投稿者はこれらのユーザに限られないため，本研究では投稿時間間隔については重複レシピ検出の手がかりとして用いないこととする．

1.3 本論文の構成

本論文の構成を以下に示す．2 章では，文書の埋め込み表現の抽出に関する研究や重複レシピの検出に関する研究など，本研究の関連研究について述べ，本研究の位置づけを示す．3 章では，本研究の提案手法である 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手法について述べる．また，材料相違数によるフィルタリング手法について示す．4 章では，5 章の比較実験で用いるデータセットおよび重複レシピペア候補のアノテーション基準について述べる．5 章では，提案手法の評価実験および実験結果に基づく考察について述べる．最後に 6 章において，本研究のまとめ，今後の課題について述べる．

第2章　関連研究

本章では，重複レシピの検出に関する研究や文書の剽窃の検出に関する研究など，本研究と関連する研究について述べ，本研究がどのような位置づけであるかを明らかにする．まず，2.1 節では，重複レシピの検出に関する研究について述べる．続いて，2.2 節で文書の剽窃検出に関する研究について述べる．以降，2.3 節，2.4 節では，2.1 節および 2.2 節で述べた関連研究を踏まえ，文書の埋め込み表現の抽出に関する研究，単語の埋め込み表現を用いて文書間の類似性を算出した研究についてそれぞれ述べる．最後に 2.5 節において，関連研究を踏まえた本研究の位置づけを示す．

2.1　重複レシピの検出に関する研究

久保ら [39] は，10 種類の料理をクエリとして，投稿型レシピサイトである「クックパッド[1]」と「楽天レシピ」を横断した重複レシピの検出を行った．この研究では，小高ら [36]，高橋ら [32] の知見（2.2 節参照）を基に，文字 3-gram 集合の Jaccard 係数を調理手順テキスト間類似度として類似したレシピペアの抽出を行い，以下の基準を用いて人手で重複レシピであるかアノテーションを行った．

(1) 料理の目的が同じもので，調理内容の一部が変更されているもの

(2) 料理の目的が異なるにも関わらず，目的に応じた調理内容に変化していないもの

久保らは料理の目的を比較する際に，レシピのタイトルと概略を使用した．アノテーションの結果，レシピサイトを横断した重複レシピの方が，単独のレシピサイトで存在する重複レシピより多いことを明らかにした．また，久保らは重複レシピにおける調理手順テキストの内容の差異の分析を行い，調理内容の一部の代替，調理内容の一部の追加，文の分割が行われていることを明らかにした．本研究では，久保らのアノテーション基準を一部踏襲しつつ，楽天レシピの事業部に対して聞き取り調査を行った結果を基に作成した新たなアノテーション基準を用いる．また，本研究ではレシピサイトを横断した

[1]クックパッド: https://cookpad.com/

重複レシピについては扱わず，楽天レシピ内に存在する重複レシピの検出について研究を行う．

著者の先行研究 [21] では，久保らと同じく文字 3-gram 集合の Jaccard 係数を調理手順テキスト間類似度として類似したレシピペアの抽出を行った．また，調理手順テキスト間の類似度に加えて，料理画像間の類似度を組み合わせて，重複レシピを判別した．料理画像間の類似度を測る際には，レシピの画像の色数を減色し，400 × 400 ピクセルの画像に変換して，各ピクセルごとの色情報を基に画像間の類似度を算出した．この研究では，重複レシピか否かを人手で評価する際に，ブラック，グレー，ホワイトからなる 3 段階の基準を設けて，アノテーションを行った．本研究では，この研究と異なり，料理画像間の類似度については考慮しない．また，この研究におけるアノテーション基準を踏襲しつつ，楽天レシピの事業部に対して聞き取り調査を行った結果を基に作成した新たなアノテーション基準（4.2 節参照）を用いる．

上記の研究では，いずれも文字 3-gram 集合間の Jaccard 係数を調理手順テキスト間類似度として重複レシピの検出を行った．しかし，著者らの先行研究 [33] において指摘した通り，ご飯とライスのように同じ材料であるにも関わらず言い換えられるものや，漢字，ひらがな，カタカナなど，複数の表記方法を持つ材料が存在する．また，本研究で扱う投稿型レシピサイトのデータは，一般ユーザによって作成されたものであるため，調理手順テキスト中に誤字や脱字が多数存在する．そのため，単に文字 3-gram を比較するだけでは，検出できない重複レシピが存在する．

以上の点を踏まえると，重複レシピの検出を行うためには，調理手順テキストの文字列的な一致度だけでなく，調理手順テキストの意味的な特徴についても考慮する必要がある．著者の先行研究 [33, 20, 34] では，意味的な特徴を考慮することを目的として，埋め込み表現を用いた重複レシピ検出手法を提案した．まず，著者の 2018 年の研究 [33] では，Sparse Composite Document Vectors [18]（2.3 節参照）で文書の埋め込み表現を作成し，重複レシピを検出する手法を提案した．また，著者の 2019 年の研究 [20, 34] では，Kusner et al. [13] が提案した WMD（2.4 節参照）を応用し，文字 3-gram の埋め込み表現に着目して重複レシピを検出する手法（Character 3-gram Mover's Distance）をそれぞれ提案した．しかし，これらの研究では，久保ら [39] が提案した文字 3-gram 集合間の Jaccard 係数を調理手順テキスト間類似度として重複レシピの検出を行う手法と同程度の検出結果となった．本研究では，重複レシピの検出を行う際に，文字列的な一致度だけでなく，意味的な特徴についても考慮することを目的として，BERT（2.3 節参照）を用いて抽出した調理手順テキストの特徴量と WMD に基づき，重複レシピを検出する

手法を提案する．

また，著者らの先行研究 [33, 34] では，材料相違数により重複レシピペア候補をフィルタリングする手法を提案しているが，その効果については定量的な分析がなされていない．そこで，本研究では材料相違数によるフィルタリングを行わない手法と比較し，材料相違数の効果の定量的な分析を行う．

2.2 文書の剽窃検出に関する研究

文書の剽窃の検出に関する研究として，小高ら [36] は文字 n-gram を用いて剽窃したレポートを検出する手法を提案している．この研究では，日本語の特徴から文字 3-gram が最も剽窃を検出するのに有効であることを示し，実験の結果，文の入れ替えや文末表現の変更などを行ってもレポートの類似性を検出できることを明らかにし，文字 3-gram を用いた剽窃検出の一定の頑健性を示した．実験では，学生が作成したレポートに基づいて，文末表現の変更，頻出単語の置換，文の出現順序の入れ替え，文の挿入を機械的に行ったレポートを作成し，提案手法で検出できるかの評価を行った．

また，高橋ら [32] は小高らの研究を応用し，Web ページの一部を剽窃して作成されたレポートを発見するシステムを提案している．実験では，Web から収集したデータを用いて作成した剽窃レポートと実際の授業で回収されたレポートを用いて実装したシステムによる剽窃文書の検出を行い，提案システムの有効性を示した．

2.1 節で示した通り，小高ら，高橋らの知見に基づき，複数の重複レシピの検出に関する研究 [39, 21, 35, 38] において，文字 3-gram 集合間の Jaccard 係数が調理手順テキスト間の類似度として用いられた．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上に，ユーザ投稿型レシピサイトに投稿されるレシピは投稿者によって使用する料理用語の表記が異なり，誤字や脱字が含まれているという特徴がある．そのため，文字の類似性のみ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．そこで本研究では，埋め込み表現を用いることにより，言い換えや書き換え，語順の入れ替えに対しても頑健な重複レシピ検出手法を提案する．なお本研究では，文字 3-gram 集合間の Jaccard 係数を用いて重複レシピの検出を行う手法を，提案手法の検証実験の際に比較手法として用いる．

2.3 文書の埋め込み表現の抽出に関する研究

2.1 節，2.2 節で述べた通り，文字列の一致度を手がかりとして重複レシピを検出した場合，検出できない重複レシピが存在する．そこで，本研究では埋め込み表現を用いて，調理手順テキスト間の意味的な一致度も考慮した重複レシピ検出手法を提案する．本節では，文書の埋め込み表現の抽出に関する研究について述べ，本研究で用いる調理手順テキストの埋め込み表現の抽出手法を示す．

Mekala et al. [18] は文書の埋め込み表現の抽出手法である Sparse Composite Document Vectors（SCDV）を提案した．SCDV は以下の手順で抽出される．

(1) 単語の埋め込み表現をクラスタリング

(2) 各単語がどのクラスタに属するかの確率を考慮して新たな単語の埋め込み表現を生成

(3) 文書中の各単語について，新たに生成された単語の埋め込み表現の各要素の総和を算出し文書の埋め込み表現を生成

(4) 得られた文書の埋め込み表現から閾値を下回る素性を 0 に変換し，疎ベクトルを生成

上記の手順で生成された疎ベクトルが SCDV として用いられる．SCDV は，文書の特徴量を抽出する他の手法 [7, 15, 16] との比較において，高い文書分類精度を達成している．著者の先行研究 [33] においても，SCDV を用いて重複レシピを検出する手法を提案した．先行研究では，最近傍探索手法である Neighborhood Graph and Tree (NGT) [27] と組み合わせることにより，文字 3-gram 集合間の Jaccard 係数と比べて高速な重複レシピ検出を実現した．その一方で，重複レシピの検出精度では，文字 3-gram 集合間の Jaccard 係数を基に重複レシピを検出する手法と同等の結果に留まった．

他にも，文書の埋め込み表現の抽出手法として，Encoder-Decoder モデル [28] を応用した Skip-thought [11] や Quick-thought [17]，Smooth Inverse Frequency [1] が提案されている．中でも，Devlin et al. [4] によって提案された事前学習言語表現モデルの BERT は，大規模コーパスによって事前学習を行うことにより，様々なタスクにおいて最高精度を達成した．Devlin et al. によると，BERT は文書の意味的な等価性を判定するタスクである Microsoft Research Paraphrase Corpus（MRPC）[5] をはじめとする，複数の自然言語処理のタスクで State of The Art（SoTA）を達成している．BERT は双方向の Transformer [30] を用いたモデルであり，単方向の Transformer を用いる OpenAI Generative Pre-trained Transformer（GPT）[2] や双方向の Long Short-Term Memory（BiLSTM）

[2] https://github.com/openai/gpt-2

[8, 25] を用いる Embeddings from Language Models（ELMo）[22] とは異なるタスクによる事前学習を行い，高い精度を達成した．BERT では，入力として入力文の先頭に [CLS] トークンを付加した単語系列を与えると，出力として単語埋め込み表現，文順序埋め込み表現，位置埋め込み表現を考慮した埋め込み表現を獲得できる．本研究では，BERT を用いて抽出した [CLS] トークンの埋め込み表現を調理手順テキストの埋め込み表現として，重複レシピの検出に用いる．

2019 年 11 月現在，公開されている日本語の BERT 事前学習モデルの主なものを表 2.1 に示す．

表 2.1: 公開されている日本語の BERT 事前学習モデル

	訓練データ	セパレータ	語彙数
柴田ら [37]	日本語 Wikipedia	Juman++ [29] & BPE [26]	32,000
森	ビジネスニュース	MeCab & NEologd	32,000
菊田[3]	日本語 Wikipedia	SentencePiece [12]	32,000
ホットリンク[4]	SNS	SentencePiece	32,000

本研究では，複数の BERT モデルを用いて重複レシピの検出を行った予備実験の結果を踏まえ，柴田らによって作成された事前学習モデルを採用する．

2.4 単語の埋め込み表現に基づき文書間の類似性を算出する研究

本節では，単語の埋め込み表現に基づき，文書間の類似性を算出する研究について述べる．Kusner et al. [13] は，単語の埋め込み表現に基づき文書間の距離を算出する手法として Word Mover's Distance (WMD) を提案している．WMD は 2 つの文書 A, B 間の距離を各文書の単語同士を対応付けることで文書 A を文書 B に置き換えるとき，対応付けのコストが最も低い場合のコストの総和と定義される．ここで，単語 i を単語 j に対応付けるコストとは，各単語の埋め込み表現間の L2 ノルムを指す．すなわち，単語 i を単語 j に対応付けるコスト $c(i, j)$ は 2.1 式で定義される．

$$c(i, j) = ||\boldsymbol{x}_i - \boldsymbol{x}_j||_2 \tag{2.1}$$

ただし，2.1 式における $\boldsymbol{x}$ は，単語ベクトルを指す．しかし，上記の式では 2 文書間の単語が 1 対 1 に対応している場合しか考慮されていない．そこで，Kusner et al. は，

D: Obama speaks in Illinois.

D': The President greets the press in Chicago.

上記の 2 文のように，文中の単語数が異なっており，単語が 1 対 1 に対応していない場合についても，以下のように Earth Mover's Distance（EMD）[24] と呼ばれる最適化問題として定式化を行い，文書間の距離を求めた．

(1) まず，文書の特徴表現として語の頻度分布を用いる．上記の *D*，*D'* の頻度分布は 2.2 式の通りである．なお，このときベクトルの各成分は，文書中の語の正規化された出現頻度となっている．

$$\begin{array}{cccccccccccc} & & \text{chicago} & & \text{greet} & & \text{illinois} & & \text{obama} & & \text{president} & & \text{press} & & \text{speak} & \\ d & = & (& 0, & 0, & 1/3, & 1/3, & 0, & 0, & 1/3 &) \\ d' & = & (& 1/4, & 1/4, & 0, & 0, & 1/4, & 1/4, & 0 &) \end{array} \tag{2.2}$$

(2) 続いて，各語に対応する成分を変換先の文書の複数の語に分配し，語の頻度分布を変換先の分布への移動を考える．この変換は 2.3 式のように示される．なお，2.3 式は，考えられる変換の一例であり，変換の中でコストの総和が最小になるものを WMD とする．

$$\mathbf{T} = \begin{pmatrix} & \text{chicago} & \text{greet} & \text{illinois} & \text{obama} & \text{president} & \text{press} & \text{speak} \\ \text{chicago} & 0, & 0, & 0, & 0, & 0, & 0, & 0 \\ \text{greet} & 0, & 0, & 0, & 0, & 0, & 0, & 0 \\ \text{illinois} & 1/4, & 0, & 0, & 0, & 0, & 1/12, & 0 \\ \text{obama} & 0, & 0, & 0, & 0, & 1/4, & 1/12, & 0 \\ \text{president} & 0, & 0, & 0, & 0, & 0, & 0, & 0 \\ \text{press} & 0, & 0, & 0, & 0, & 0, & 0, & 0 \\ \text{speak} & 0, & 1/4, & 0, & 0, & 0, & 1/12, & 0 \end{pmatrix} \tag{2.3}$$

なお，上記の対応付きコストはそれぞれの語同士の対応コストに，分配量に応じた重みを掛けた値の総和として計算できる（2.4 式）．

$$\sum_{i,j} \mathbf{T}_{ij}\, c(i,j) \tag{2.4}$$

ここで，2.3 式で示される行列 $\mathbf{T}$ の各行ベクトルは，元の文書中の各語に対応し，各成分は元の文書の語分布におけるその語の成分を，変換先の文書の各語にどのように分配するかを表す． よって行の成分の総和は，元の文書の語分布における各語の成分と一

致する（2.5 式）.

$$\sum_{j} \mathbf{T}_{ij} = d_i \tag{2.5}$$

同様に，行列の各列ベクトルの成分の総和は，変換先の文書の語分布に各語の成分と一致する（2.6 式）.

$$\sum_{i} \mathbf{T}_{ij} = d'_j \tag{2.6}$$

WMD では，コストが最小となる変換のコストを文書間の距離とする．そこで，上式を制約条件とする 2.7 式，2.8 式，2.9 式で示される最適化問題を解くことで文書間の距離が求められる.

$$\min_{\mathbf{T} \geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij}\, c(i,j) \tag{2.7}$$

$$\text{subject to:} \sum_{j=1}^{n} \mathbf{T}_{ij} = d_i \ \forall i \in \{1, ..., n\} \tag{2.8}$$

$$\sum_{i=1}^{n} \mathbf{T}_{ij} = d'_j \ \forall j \in \{1, ..., n\} \tag{2.9}$$

Kusner et al. は実験において，類似した文書を抽出する実験を行い，多くのデータセットにおいて提案手法が比較手法を上回る結果を残した．とりわけ，Twitter のような短文からなるデータセットに対して，より良い結果が得られている．WMD は実験で示されているように精度が高く，直観的な理解に優れている．また，埋め込み表現を用いることから言い換えや書き換えにも強く，単語同士の対応付けを行うことから語順の入れ替えに対しても頑健性を持つ．また，パラメータチューニングが不要であるといった優れた特徴を持つ一方で，計算量が膨大[5]であるという課題が存在する．重複レシピの検出は，レシピサイトのサービスの品質に直結するため，正確性が求められる一方で，1 日に多くのレシピが投稿されるため，検出の速度も要求される．そのため，WMD を重複レシピの検出に用いることは実際のサービスの現場では難しい．そこで本研究では，調理手順テキストの埋め込み表現間の距離に基づいて，WMD の算出対象とするレシピペアを絞り込むことにより，重複レシピの検出速度についても重視しつつ，重複レシピの検出精度を高めることを目指す.

[5] 2 文書の異なり単語数を w_c と置くと，$O({w_c}^3 \log w_c)$

2.5 本研究の位置づけ

本研究では，投稿型レシピサイト上に存在する重複レシピの判別を目的とする．ユーザ投稿型コンテンツには，誤字や脱字などが含まれていることに加え，言い換えや書き換えを行った上で重複レシピを投稿するユーザが存在することから，単に文字列の一致度を判断基準として，重複レシピを検出することは難しい．そこで，本研究では文書の埋め込み表現に着目し，誤字や脱字などが含まれているユーザ投稿型レシピに対して頑健な重複レシピ検出手法を提案する．また，重複レシピの検出における材料相違数によるフィルタリングの有効性について検証する．

2.3 節で示した通り，文書の埋め込み表現を抽出するための手法として，複数の手法が提案されている．本研究では，複数の自然言語処理タスクにおいて高い精度を残している BERT を用いて文書埋め込み表現を抽出する．ただし，文書の埋め込み表現を用いて，重複レシピを抽出した場合，調理手順テキスト中の単語のアラインメントが考慮されない．そこで，文書の埋め込み表現を用いて，類似したレシピを抽出した上で，WMD を用いて重複レシピペア候補をリランキングすることによって，重複レシピの検出を行う手法を提案する．

第3章　提案手法

本章では，本研究の提案手法である 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手法について述べる．はじめに，3.1 節において，提案手法の概要を示す．続いて，3.2 節では，調理手順テキストの埋め込み表現間の距離に基づく重複レシピペア候補のランキング，3.3 節では，材料相違数による重複レシピペア候補のフィルタリング，3.4 節では，WMD に基づく重複レシピペア候補のリランキングについてそれぞれ述べる．

3.1　提案手法の概要

本研究では，(1) BERT を利用して抽出した調理手順テキストの埋め込み表現間の距離，(2) 重複レシピペア候補の材料相違数によるフィルタリング，(3) 単語埋め込み間の対応に着目して文書間の距離を算出する WMD に基づき，重複レシピを検出する手法を提案する．提案手法のねらいとして，以下の 3 点が挙げられる．

- 文書の意味的な等価性を判定するタスクである Microsoft Research Paraphrase Corpus（MRPC）[5] をはじめとする，複数の自然言語処理のタスクで最高精度を達成した BERT を用いて調理手順テキストの埋め込み表現を抽出することで，調理手順テキストの意味についても捉えた重複レシピの検出を行う
- 材料相違数によるフィルタリングを行うことで，調理手順テキストは共通しているものの，材料がまったく異なる料理について記述したレシピペアを重複レシピとして誤検出することを防ぐ
- WMD に基づいて，重複レシピペア候補のリランキングを行うことで，BERT を用いて調理手順テキストの埋め込み表現を抽出した場合に調理手順テキスト中の単語間の意味的な対応が考慮されない点を補う

本研究における提案手法の概要を図 3.1 に示す．本研究の提案手法は以下の 3 段階から構成される．

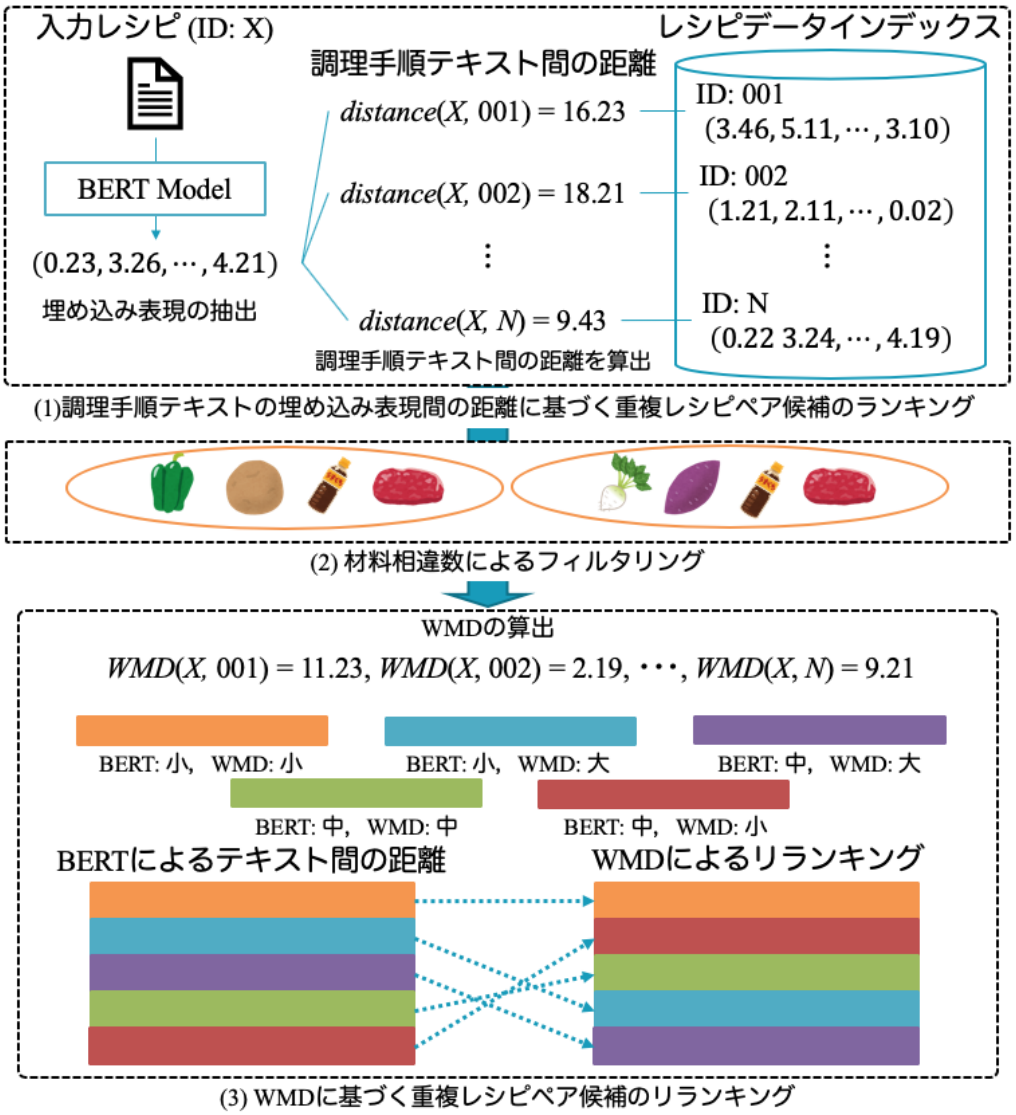

図 3.1: 提案手法の概要図

（1）調理手順テキストの埋め込み表現間の距離に基づく重複レシピペア候補のランキング（3.2 節）

（2）材料相違数に基づく重複レシピペア候補のフィルタリング（3.3 節）

（3）WMD に基づく重複レシピペア候補のリランキング（3.4 節）

以降，3.2 節，3.3 節，3.4 節において，本研究の提案手法の各ステップについて詳細に述べる．

3.2 調理手順テキストの埋め込み表現間の距離に基づく重複レシピペア候補のランキング

本節では，調理手順テキストの埋め込み表現間の距離および材料相違数に基づいて重複レシピペア候補のランキングを行う手法について述べる．本ステップは，大きく BERT を用いて調理手順テキストの埋め込み表現の抽出する部分，調理手順テキストの埋め込み表現間の距離を算出する部分の 2 つに分かれる．

BERT を用いた調理手順テキストの埋め込み表現の抽出: 本研究では，調理手順テキストの埋め込み表現の抽出に柴田ら [37] によって公開されている日本語の学習済み BERT モデルを使用する．柴田らのモデルの詳細を以下に示す．

- 入力テキスト: 日本語 Wikipedia の全データ（約 1,800 万文，半角を全角に変換）
- 入力テキストに Juman++ (v2.0.0-rc2) [29] で形態素解析を行った上で，Bite Pair Encoding (BPE) [26] を適用し subword [2] に分割
- モデルの構成: 双方向 Transformer 12 層，隠れ状態ベクトル 768 次元，Attention のヘッド数 8 個
- 学習エポック数: 30 回
- 語彙数: 32,000 語（形態素，subword を含む）
- 最大入力長: 128

2 章で述べたとおり，BERT の日本語の学習済みモデルが，菊田やホットリンク社によって公開されている．柴田らのモデルは日本語 Wikipedia をコーパスとして用いて学習されたものであり，ホットリンク社のモデルは 2017 年および 2018 年に投稿された日本語ツイートの一部を用いて学習されたものである．文書の単語単位への分割には，柴田らのモデルは Juman++および BPE を，ホットリンク社のモデルは SentencePiece モデルを用いている．予備実験として，柴田らのモデルとホットリンク社のモデルを用いて，重複レシピ検出精度の比較を行った．予備実験では，5 章で述べる手順で両モデルを用いて調理手順テキストの埋め込み表現を抽出した上で，重複レシピペア候補の抽出を行い，アノテーションを行うことで，重複レシピ検出精度を比較した．その結果，ホットリンク社のモデルは柴田らのモデルより，重複レシピの検出数が 1 割程度減少した．4 章で示すアノテーション基準における「重複-その他」にあたるレシピペアは，柴田らの

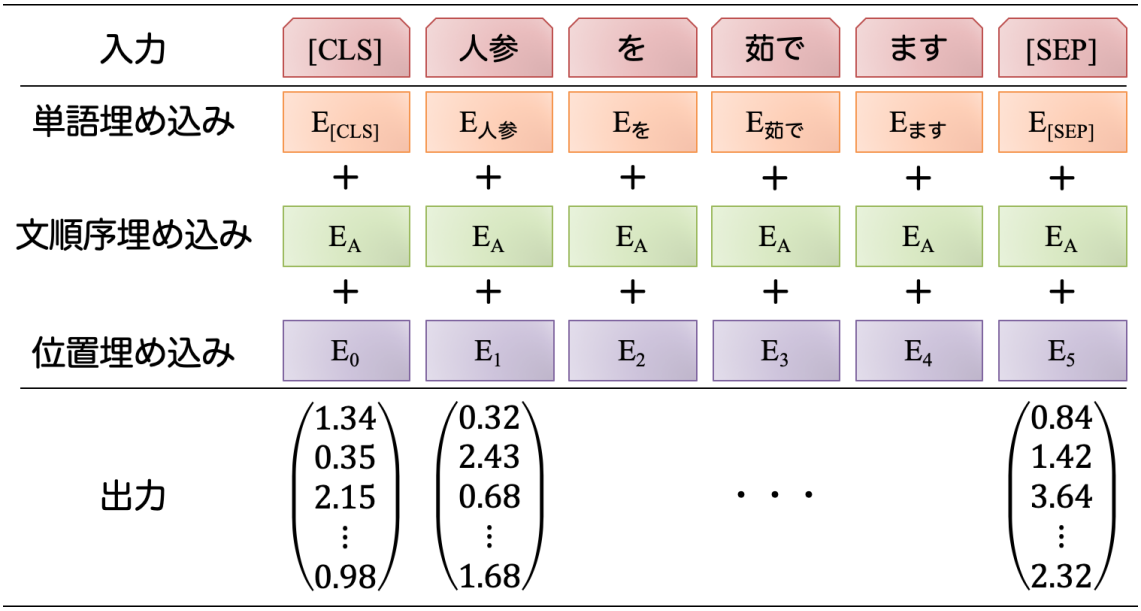

図 3.2: BERT における調理手順テキストの埋め込み表現の抽出方法

モデルでは重複レシピペア候補に 1 件も含まれていなかったが，ホットリンク社のモデルでは数件検出された．この結果より，Juman++によって，文を分割した上で BPE を適用し，文を分割する柴田らのモデルの方が，レシピの意味的な特徴を抽出するのに適していると考え，柴田らのモデルを採用する．

調理手順テキストの埋め込み表現を抽出する際には，まず調理手順テキストを全角に統一した上で，Juman++を用いて，調理手順テキストを形態素単位へと分割する．そして，形態素単位へと分割した調理手順テキストの先頭に [CLS] トークンを付け加えた系列を BERT モデルへの入力として与える．なお，柴田らのモデルでは，モデル内部で各形態素について BPE を適用し，Subword 単位での分割を行う．

モデルからの出力として，文中の各語の埋め込み表現が得られる（図 3.2）．図 3.2 に示す通り，BERT では入力された各語の単語埋め込み，文順序埋め込み，位置埋め込みを用いることによって，出力の埋め込み表現を抽出する．本研究では，出力される系列の先頭にある [CLS] トークンの埋め込み表現を調理手順テキストの埋め込み表現として利用する．

なお，実験では BERT の 1 層から 12 層までの各層において [CLS] トークンから得られた 768 次元のベクトルからなる調理手順テキストの埋め込み表現を抽出する．これにより，重複レシピの検出タスクにおいて，BERT のどの層から出力される埋め込み表現が有効であるかを調査する．

調理手順テキストの埋め込み表現間の距離の算出: 続いて，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を Johnson et al. [10] によって提案された最近傍探索ライブラリである faiss[1]を用いて算出する．faiss では，ベクトル集合をインデックスに格納することで，クエリを入力として与えた際に高速な計算が可能になる．重複レシピの検出は，検出精度に加えて，実時間での検出が必要とされるタスクである．そのため，本研究では，faiss を調理手順テキスト間の距離の算出に用いることで処理の高速化を図る．

本研究において調理手順テキスト間の距離を算出する際には，予めオリジナルレシピ候補群の調理手順テキストの埋め込み表現をレシピデータインデックスに格納しておき，入力として重複レシピ候補の調理手順テキストの埋め込み表現を与える．そして，重複レシピ候補とオリジナルレシピ候補群の各レシピの間の距離を総当たりで算出する．なお，本研究では調理手順テキストの埋め込み表現間の L2 ノルムを調理手順テキスト間の距離とする．すなわち，調理手順テキスト x と調理手順テキスト y から抽出した調理手順テキストの埋め込み表現 $x_{embedding}$ と $y_{embedding}$ の間の距離は 3.1 式で定義できる．

$$distance(x, y) = ||\boldsymbol{x}_{embedding} - \boldsymbol{y}_{embedding}||_2 \tag{3.1}$$

3.3 材料相違数に基づく重複レシピペア候補のフィルタリング

本節では，材料相違数に基づいて，3.2 節で抽出した重複レシピペア候補をフィルタリングする方法を述べる．レシピの中には，異なる料理について記述したレシピであっても，重複した調理手順テキストになるものがある．例えば，フルーツヨーグルトとサラダは，料理としては異なっている．しかし，調理手順テキストに単に「材料をすべて混ぜて完成」と記述されているレシピが存在する．こうしたレシピの取り扱いについて，楽天レシピの事業部に聞き取り調査を行った結果，質の低いレシピではあるものの，重複レシピとしての削除をするまでには至らないとの回答を得た．この基準に照らして考えると，単に調理手順テキスト間の距離のみを重複レシピ検出の手がかりとして用いた場合，調理手順テキストは共通しているが，材料がまったく異なる料理について記述したレシピペアを重複レシピとして誤検出する可能性がある．そのため，本研究では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

[1]facebookresearch/faiss（GitHub）: https://github.com/facebookresearch/faiss

検出防止を目的として，材料相違数に基づく，重複レシピペア候補のフィルタリングを行う．

数多の食材の中には，同じ食材であるにも関わらず言い換えが可能な食材が存在する．また,「塩こしょう」「塩コショウ」「塩胡椒」のように，日本語で材料について記述した場合，複数の表記方法がある材料がある．そのため，単に文字列のみを比較しただけでは，材料相違数の正確な算出が行えない．

そこで，本研究ではより高い精度での材料相違数の算出を目的として，先行研究 [33, 34] に基づき，以下の手順によってオリジナルレシピ候補の材料集合と重複レシピ候補の材料集合間の材料相違数を算出する．

（1）両レシピの材料リストから記号を取り除き，括弧内等の文字は削除する．(例: 「●にんにく」となっていた場合,「●」を取り除き,「にんにく」とする．また，りんご (青森県産) となっていた場合，括弧を削除することに加えて，括弧内の文字を削除し,「りんご」とする．)

（2）材料名をすべて全角カタカナ表記に統一し，文字列が両レシピ間で完全一致する材料を両レシピの材料リストから削除する．(例: リンゴ (りんごから変換) とリンゴ (林檎から変換)) なお，材料名を全角カタカナに変換する際には，pykakasi[2]を用いた．

（3）重複レシピ候補の各材料について，レシピデータを用いて学習した単語の分散表現をもとに，類似単語を検索する．類似単語の検索には gensim[3]を用いる．類似単語の検索結果，上位 5 件にオリジナルレシピ候補の材料が含まれていた場合，同一の材料とみなして，クエリとした材料および類似単語として抽出された材料を材料リストから削除する．(例: 米とライス，お米と米)

（4）両レシピの材料リスト中の要素の合計数を材料相違数とする．

なお，本研究では，先行研究 [34] の手法を踏襲し，材料相違数が 2 以下の場合に重複レシピペア候補として抽出する．予備実験として，材料相違数の閾値を 2 とすることの妥当性を検証するために，材料相違数を 3 以上とした場合の重複レシピ検出数を確認した．その結果，材料相違数を 3 以上と変化させても，非重複レシピが増加するのみで，調理手順テキスト間の類似度（距離）上位のレシピに重複レシピが出現しなかった．よって，材料相違数の閾値を 2 とするのは妥当といえる．

[2]https://github.com/miurahr/pykakasi
[3]gensim: https://radimrehurek.com/gensim/

3.4 WMD に基づく重複レシピペア候補のリランキング

本節では，WMD を用いて重複レシピペア候補のリランキングを行う手法について述べる．本ステップは，重複レシピペア候補の WMD の算出を行う部分，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および WMD を用いた重複レシピペア候補のリランキングを行う部分の 2 つに分かれる．

重複レシピペア候補の WMD の算出: まず，重複レシピペア候補中の各レシピペア間の WMD を算出する．WMD は，2 つの文書 A, B 間の距離を各文書の単語同士を対応付けることで文書 A を文書 B に置き換える際に，対応付けのコストが最も低い場合のコストの総和と定義されている．

例えば，以下の 3 文における

D_0 人参を切ります。
D_1 にんじんを切る。
D_2 じゃがいもを茹でる。

D_1 から D_0，D_2 から D_0 への置き換えは図 3.3 のように算出される．図 3.3 における D_0

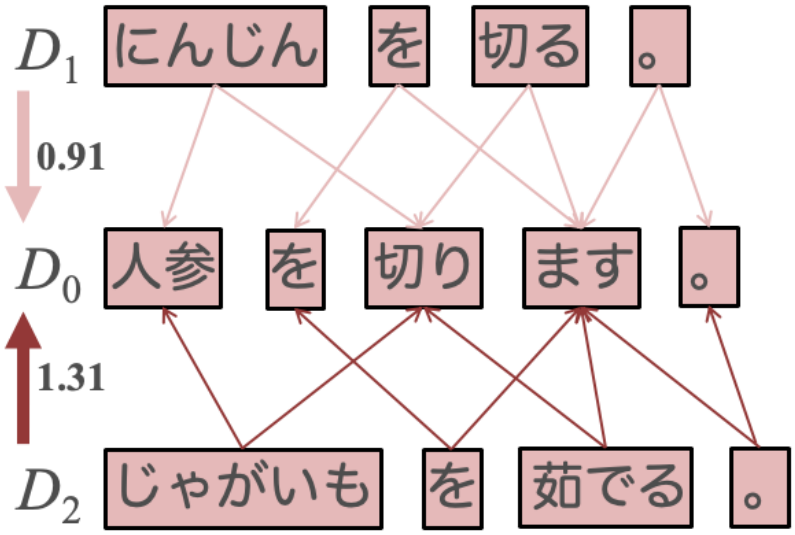

図 3.3: Word Mover's Distance の計算における単語間の対応の例

と D_1 の間の文書間の距離は 0.91，D_0 と D_2 の間の文書間の距離は 1.31 であり，D_0 と D_1 の方が D_0 と D_2 より距離が近い，すなわち類似しているといえる．

本研究では，WMD の算出対象とする重複レシピペア候補を，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材料相違数を基づいて選択する．これに

より，WMD の算出回数を減少させることができ，重複レシピ検出の処理の高速化が図れる．

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および WMD を用いた重複レシピペア候補のリランキング: 続いて，日本語の学習済み BERT モデル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 WMD に基づいて，重複レシピペア候補をリランキングする．前述の通り，文書の特徴量として，BERT を用いて調理手順テキストの埋め込み表現を抽出した場合，調理手順テキスト中の単語間の意味的な対応が考慮されない．そのため，WMD を用いて重複レシピペア候補のリランキングを行うことで，単語間の対応にも着目し，より頑健に重複レシピの検出を行うことを目指す．

BERT による調理手順テキスト間の距離および WMD では，どちらとも距離の最短は 0 となるものの，最長の距離についてはそれぞれの手法で異なる．そこで，3.2 式を用いて，どちらの距離指標でも距離が 0 から 1 の間を取るよう 2 つの距離指標を正規化する．

$$Y = \frac{X - x_{min}}{x_{max} - x_{min}} \tag{3.2}$$

なお，3.2 式において，X は正規化を行う前の 2 文書間の距離を指し，Y は正規化後の 2 文書間の距離を指す．また，x_{min} は，リランキング対象とする重複レシピペア候補における最短の調理手順テキスト間の距離を指し，x_{max} は，リランキング対象とする重複レシピペア候補における最長の調理手順テキスト間の距離を指す．

第4章　実験データおよび重複レシピペア候補のアノテーション基準

本章では，実験に用いるデータセットおよび重複レシピペア候補のアノテーション基準，アノテーション基準の妥当性について検証した結果を述べる．4.1 節において，実験に用いるデータセットの概要について述べ，4.2 節において楽天レシピの投稿規約および楽天レシピの事業部に対する聞き取り調査に基づき作成した重複レシピペア候補のアノテーション基準について述べる．最後に，4.3 節において，4.2 節で示したアノテーション基準を用いて，4 名のアノテータによってレシピペアに対してラベリングを行い，アノテーション基準の妥当性を検証した結果を述べる．

4.1　実験に用いるデータセット

実験では，楽天株式会社より提供を受けた楽天レシピのデータセットを用いる．本データセットは，2010 年 6 月 30 日から 2018 年 9 月 13 日までの間に楽天レシピに投稿された 1,668,270 件のレシピについて，投稿者 ID，レシピ ID，レシピの投稿時間，調理手順テキストなどの情報から構成されている．

実験では，レシピデータインデックスに格納するデータ，すなわちオリジナルレシピ候補群のデータとして，2010 年 6 月 30 日から 2018 年 8 月 31 日までの間に投稿された 1,661,997 件のレシピを用いる．また，これらのレシピデータを用いて，材料相違数および WMD を算出する際に必要となる単語埋め込み表現モデル（学習アルゴリズム: Skip-Gram model with Negative Sampling （SGNS），埋め込み表現の次元数: 100 次元，窓幅: 15）[19] についても作成する．テストデータ，すなわち実験において重複レシピ候補群として扱うデータとしては，2018 年 9 月 1 日から 2018 年 9 月 13 日までの間に投稿された 6,273 件のレシピを用いる．

すなわち，実験においては 2018 年 8 月 31 日までに投稿されたレシピ群が楽天レシピに掲載されている状況で，2018 年 9 月に新たにレシピ群が投稿される状況を想定しており，2018 年 9 月に投稿されたレシピ群の中から重複レシピを検出する．2018 年 8 月以前

に投稿された重複レシピについては，著者の先行研究の成果を基に楽天レシピに報告を行い，複数の重複レシピが削除されている．そのため，本研究では，2018 年 9 月に投稿されたレシピを重複レシピ候補とした実験を行う．

4.2 重複レシピペア候補のアノテーション基準

本節では，重複レシピペア候補として抽出されたレシピペアのアノテーションを行う際の基準について述べる．なお，本基準は楽天レシピの事業部に対する聞き取り調査の結果および楽天レシピが定めるレシピの投稿規約[1]に基づいて作成した．重複レシピ検出精度の評価を目的として，レシピペアのアノテーションを行う際には，以下の 4 段階のアノテーション基準に基づいてアノテーションを行う．アノテーション作業では，基準中のいずれか 1 つの条件を満たしていれば，該当するタグを付与し，レシピの構成要素中に重複した要素が全く含まれていない場合，非重複-その他とラベリングする．

なお，重複と非重複の 2 段階ではなく，非重複の中でも細分化した理由として，楽天レシピの事業部における聞き取り調査において，重複レシピ，すなわちレシピサイトの投稿規約に違反しており，削除をする必要があるレシピの他にも，投稿規約の違反とは言えないが，他のレシピの調理手順テキストと完全に一致しており，1 つの材料を入れ替えたようなレシピの存在が明かされたことがあげられる．このようなレシピは，削除を行うことはできないまでも，アレンジレシピとして 1 つにまとめあげることによって，表示される検索結果の多様性を維持することができる．そのため，非重複の中でもアノテーション基準を調理手順テキストおよび材料の重複度合いによって 3 段階に細分化した．

- 重複
  - 材料が完全に一致しており，調理手順テキストが文末表現等を除き一致しているレシピペア
  - 材料に主要でない材料が 1 つ程度追加されているものの，調理手順テキストには材料の変化に応じた書き換えが見られないレシピペア
- 非重複-A（表 4.1）
  - 同じ料理で，材料について共通している部分があり，調理手順テキストについても共通している部分があるレシピペア

[1]楽天レシピ投稿ガイドライン（楽天レシピ）：https://recipe.rakuten.co.jp/rules/

表 4.1: 非重複-A レシピペアの例

レシピ A	タイトル	バナナヨーグルト
	使用材料	・ヨーグルト・大麦グラノーラ・バナナ ・はちみつ
	調理手順テキスト	（1）バナナを輪切りにする （2）器にヨーグルト、大麦グラノーラ、バナナ、はちみつをかけて！
レシピ B	タイトル	きなことバナナヨーグルト
	使用材料	・プレーンヨーグルト・バナナ・きなこ ・はちみつ
	調理手順テキスト	（1）バナナを 6 ミリの輪切りにする。 （2）ヨーグルト、バナナ、きなこ、はちみつの順に器に盛る。

  - 主要でない材料が異なっているが，調理手順テキストに共通している部分があるレシピペア
  - 調理手順テキストに記載されている調理器具や調理時間等が異なるが，その他の調理手順テキストに共通している部分があるレシピペア

表 4.1 に示したレシピペアは，使用材料に相違点があり，調理手順テキストについても，バナナの調理方法に差異（一方では輪切りの方法に 6 ミリと記載されている）が見られる．しかし，レシピペアにおける異なる使用材料である「大麦グラノーラ」と「きなこ」に関する調理の詳細が記述されておらず，全体を通してみると「バナナを切って，材料を器に盛る」という共通の構成になっている．よって，このレシピペアにおける関係性は「同じ料理で，材料について共通している部分があり，調理手順テキストについても共通しているレシピペア」に該当し，非重複-A と判断できる．

- 非重複-B（表 4.2）
  - 主要な材料が異なっているが，異なる食材についての処理を除けば，調理手順テキストに共通している部分があるレシピペア
  - 異なる料理だが，調理手順，材料に共通している部分があるレシピペア

表 4.2 に示したレシピペアは，使用材料に相違点があり，調理手順テキストについ

表 4.2: 非重複-B レシピペアの例

レシピ C	タイトル	レーズンくるみパン
	使用材料	・かぼちゃ・たまねぎ・ベーコン・小麦粉 ・粉チーズ・塩こしょう・卵・パン粉・油
	調理手順テキスト	（1）☆印以外を HB に入れて生地を作る。 （2）取り出してくるみとレーズンを手でもみこみ一次発酵する。40 分で倍に! （3）10 分割して 10 分ベンチタイム。 （4）丸め直して、二次発酵 30 分。 （5）170 度で 9 分焼成
レシピ D	タイトル	ゴマ入り食パン
	使用材料	・強力粉・薄力粉・砂糖・塩・水・ごま ・インスタントドライイースト・すりごま ・マーガリン
	調理手順テキスト	（1）全て材料を hb に入れて生地を作り一次発酵。40 分で倍ほどに （2）ガス抜きして 2 分割して 10 分のベンチタイム （3）ベンチタイム後、丸め直し型に入れて二次発酵で 30 分 （4）形のこれくらいまでなったらオッケー （5）ふたをして 180 度で 30 分焼成

ても，レシピ C では，レーズンの調理方法に関する詳細な記載がされている．一方，使用材料のうち，強力粉，薄力粉，砂糖，塩，インスタントドライイースト，水，マーガリンが両レシピで共通しており，調理手順テキストについても言い回しが変更されている点が見受けられるものの全体を通してみると調理手順テキスト中の表現や構成等に関して共通している点が多い．よって，このレシピペアにおける関係性は「主要な材料が異なっているが，異なる食材についての処理を除けば，調理手順テキストに共通している部分があるレシピペア」に該当し，非重複-B と判断できる．

- 非重複-その他
  - 上記，重複，非重複-A，非重複-B のどれにも当てはまらないもの

なお，アノテーション基準中の**文末表現**とは，調理手順テキストにおける文末の言い回しのことを指す．例えば,「～を焼く」と「～を焼きます」という 2 文があった場合，言い回しは異なっているが，重複しているとみなす．他にも，一方のレシピの調理手順テキストの末尾に「できあがり」と付け加えられている場合についても，他の部分の内容が重複している場合は，重複しているとみなす．

また，**主要な材料**とは，肉や魚，野菜といった当該のレシピを構成する上での核となる材料のことを指す．**主要でない材料**とは，七味唐辛子やたれといった，料理の上に乗せたり，混ぜ合わせたりする材料のことを指す．

材料の変化に応じた書き換えとは，一方のレシピから付け加えられた材料があった場合に調理手順にその材料に応じた書き換えを行っている状態を指す．例えば，一方のレシピから「七味」が追加されている場合に，調理手順テキスト中に七味に関する調理手順の追記が明確になされていれば，材料の変化に応じた書き換えがされているとみなす．一方で，調理手順テキスト中に七味に関する言及がないものや「材料 A と材料 B を混ぜます」を「材料 A と材料 B，七味を混ぜます」のようにただ追加した材料について追記しているだけのものは材料に応じた書き換えがされているとはみなさない．

4.3 アノテーション基準の妥当性の検証

本節では，アノテーション結果に基づき，アノテーション基準の妥当性を評価した結果について述べる．実験では，重複レシピ検出手法の評価を行うにあたり，20 代の学生 3 名（男性 1 名，女性 2 名）を実験参加者として雇用し，著者を加えた 4 名のアノテータによって，5 章で述べるアノテーション対象のレシピペアから共通する 500 件のレシピペアに対してアノテーションを行った．

本研究では，アノテーション基準の妥当性を，3 名以上のアノテータ間の回答の一致度を示す指標である Fleiss [6] の κ 係数ならびに 2 名のアノテータ間の回答の一致度を示す指標である Cohen [3] の κ 係数を用いて評価する．まず，4 人のアノテータの回答の一致度を Fleiss の κ 係数で評価した結果，0.654（Substantial Agreement [14]）となった．

続いて，表 4.3 に Cohen の κ 係数を算出した結果を示す．表 4.3 より，2 名のアノテータ間の回答の一致度を示す Cohen の κ 係数は，いずれの回答者の組み合わせにおいても 0.8 以上（Almost Perfect Agreement [14]）であった．

上記の結果から，本研究で用いるアノテーション基準の妥当性が確認された．また，Cohen の κ 係数がいずれのアノテータの組み合わせにおいても 0.8 を上回ったことから，

表 4.3: アノテータ間の回答一致度 (Cohen の κ 係数)

アノテータ ID	B	C	D
A	.942	.956	.934
B	-	.908	.886
C	-	-	.902

アノテータによってアノテーション結果に大きな差異が生まれないことを確認できた．以上の点を踏まえて，他のアノテーション対象となるレシピペアについては，著者一人がアノテーション作業を担当した．なお，本研究では重複レシピの存在が投稿型レシピサイトに悪影響を及ぼす可能性がある性質を考慮し，いずれかのアノテータが「重複」とラベリングしたレシピペアについて重複レシピペアとみなす．

第5章　比較実験

本章では，本研究の提案手法である 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検出手法の有効性の検証結果および重複レシピの検出における材料相違数によるフィルタリングの有効性の検証を行った結果を示す．はじめに 5.1 節において，本実験の目的を示す．以下，5.2 節において実験方法について，5.3 節において提案手法の実現方法について，5.4 節において実験における比較手法について，5.5 節において実験結果についてそれぞれ述べる．最後に 5.6 節において，実験結果を踏まえた考察について述べる．

5.1　実験の目的

本実験では，重複レシピの検出実験を行うことで，以下の 3 点について明らかにすることを目的とする．

- 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有効性
- 材料相違数によるレシピペアのフィルタリング手法の有効性
- WMD を用いた重複レシピペア候補のリランキング手法の有効性

これらの検証を目的として，本実験では（1）提案手法と過去の重複レシピ検出に関する研究で成果を残している 3-gram 手法など複数の比較手法との比較,（2）材料相違数によるフィルタリングを行う手法と行わない手法の比較,（3）WMD によるリランキングを行う手法と行わない手法の比較を行う．

実験では，4.1 節で述べたように，楽天レシピのデータセットをレシピの投稿日を基準として，オリジナルレシピ候補群とするデータと重複レシピ候補群とするデータに分割する．すなわち，2018 年 9 月に投稿されたレシピ群の中から重複レシピを抽出するケースを想定して実験を実施する．

なお，本研究では以下の 2 つの手法を提案手法として実験を行う．

- BERT により抽出した調理手順テキストの埋め込み表現間の距離と材料相違数に基づく重複レシピペア候補のフィルタリングにより重複レシピを検出する手法（**nd_BERT**（near-duplicate BERT））
- BERT により抽出した調理手順テキストの埋め込み表現間の距離と材料相違数に基づく重複レシピペア候補のフィルタリングおよび WMD により重複レシピを検出する手法[1]（**nd_BERT-WMD**）

5.2 実験方法

実験では，5.1 節で述べた 2 つの提案手法と 5.4 節で示す 3 つの比較手法を用いて，重複レシピ候補とオリジナルレシピ候補との類似性に基づき重複レシピペア候補を抽出する．そして，4.2 節において述べたアノテーション基準に則って，重複レシピペア候補に対してアノテーションを行い，重複レシピ検出精度を評価する．本実験では，各手法について，調理手順テキスト間の距離（類似度）に基づき，重複レシピペア候補のランキングを行った上で，調理手順テキスト間の距離（類似度）上位 300 件の重複レシピペア候補をアノテーション対象のレシピペアとする．なお，距離（類似度）が同じレシピペアが複数あった場合，調理手順テキスト中の文字列がより長く一致している方が悪質性が高いと考え，一致している文字数の長さを基にアノテーションの対象とするレシピペアを決定する．

評価を行う際には，アノテーション結果を基に，上位 50 件，100 件，200 件，300 件における重複レシピペアの検出結果を比較する．ここで，上位 300 件までを評価対象とするのは，上位 400 件，上位 500 件の重複レシピペア候補にアノテーションを行った場合でも，重複レシピペア数が増加しなかったためである．また，本研究では，重複レシピの存在が投稿型レシピサイトに悪影響を及ぼす可能性がある性質を考慮し，いずれかのアノテータが「重複」とラベリングしたレシピペアについて重複レシピペアとみなす．なお，提案手法の 1 つである nd_BERT と 3 つの比較手法については，材料相違数によるフィルタリングの有効性の検証を目的として，フィルタリングを行う場合と行わない場合でそれぞれ評価を行う．

[1]なお，本手法では調理手順テキストの埋め込み表現間の距離と WMD の積を調理手順テキスト間の距離とする

5.3 提案手法の実現方法

本節では，実験における提案手法の実現方法を述べる．単語の埋め込み表現間の距離と WMD に基づいて重複レシピを検出する手法である nd_BERT-WMD は，以下に示す 6 つのステップから構成される．ただし，単語の埋め込み表現間の距離のみに基づき重複レシピを検出する手法である nd_BERT は，以下のステップのうち（1）〜（4）のみで重複レシピを検出する．

（1）**BERT を用いた調理手順テキストの埋め込み表現の抽出**

BERT を用いた調理手順テキストの埋め込み表現の抽出を，3.2 節で述べた方法に基づいて行う．抽出の対象とするレシピペアは，実験で使用する全レシピである．

調理手順テキストの埋め込み表現を抽出する際には，対象の各レシピを Juman++ [29] で形態素ごとに分割した後に，BERT の学習済みモデルに入力として与え，モデル内部で Subword 単位に分割して，調理手順テキストに含まれる各語の埋め込み表現を抽出する．提案手法では，抽出された埋め込み表現のうち，[CLS] トークンに対応する 768 次元の埋め込み表現を抽出する．なお，実験では，12 層から構成される BERT モデルのどの層から抽出した調理手順テキストの埋め込み表現が重複レシピの検出タスクに有効であるか調査することを目的として，1 層〜12 層の全ての層から埋め込み表現を抽出する．

（2）**オリジナルレシピ候補群のレシピデータインデックスへの格納**

抽出した調理手順テキストの埋め込み表現のうち，オリジナルレシピ候補群から抽出した埋め込み表現をレシピデータインデックスへ格納する．なお，レシピデータインデックスとは，オリジナルレシピ候補群の埋め込み表現が格納されているインデックスを指す．この処理は，faiss を用いて，調理手順テキスト間の距離を高速に計算することを目的として行う．

（3）**調理手順テキスト間の距離の算出**

重複レシピ候補群中の各レシピをクエリとして，レシピデータインデックスに格納したレシピとの距離を総当たりで算出する．なお，調理手順テキストの埋め込み表現間の距離として，前述の通り調理手順テキストの埋め込み表現間の L2 ノルムを用いる．

実験では，以降の処理の高速化を目的として，重複レシピ候補群の各レシピに対して，調理手順テキスト間の距離上位 500 件のオリジナルレシピ候補を抽出する．

すなわち，手法ごとに，6,273 レシピ × 上位 500 件の計 3,136,500 件を重複レシピペア候補として抽出する．

（4）**材料相違数に基づく重複レシピペア候補のフィルタリング**
前述の手順で抽出した重複レシピペア候補を対象として，材料相違数に基づき，重複レシピペア候補のフィルタリングを行う．なお，上述した通り，本研究では，材料相違数が 2 以下の場合に重複レシピペア候補として抽出する．nd_BERT では，本手順におけるフィルタリング結果に基づき，アノテーション対象レシピを抽出する．また，本研究では，材料相違数の有効性の調査を目的として，材料相違数によるフィルタリングを行わない手法を比較手法として重複レシピの検出を行う．

（5）**重複レシピペア候補の WMD の算出**
材料相違数に基づくフィルタリングを行い抽出した重複レシピペア候補の WMD を算出する．なお，実験では，処理の高速化を目的として，材料相違数でフィルタリングを行った重複レシピペア候補の中から，調理手順テキスト間の距離上位 1,500 件の重複レシピペア候補を抽出し，それらを対象として WMD の算出を行う．

（6）**重複レシピペア候補のリランキングによるアノテーション対象レシピの抽出**
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および WMD に基づいて，重複レシピペア候補のリランキングを行いアノテーション対象レシピを抽出する（nd_BERT-WMD）．

5.4 比較手法

本実験では，3 つの比較手法との比較を行い，提案手法の有効性を検証する．以下に比較手法の概要を示す．

WMD のみを手がかりとして重複レシピを検出する手法（WMD）： 本比較手法では，WMD に基づき，重複レシピペア候補を抽出する．まず，重複レシピ候補群の各レシピをクエリとして，オリジナルレシピ候補群の各レシピに対して，総当たりで WMD を算出する．そして，WMD に基づき重複レシピペア候補を抽出する．なお，実験では材料相違数によるフィルタリングを行った上で重複レシピペア候補を抽出する手法との比較を行う．

文字3-gram集合のJaccard係数を手がかりとして重複レシピを検出する手法（3-gram）： 本比較手法では，文字3-gram集合のJaccard係数に基づき，重複レシピペア候補を抽出する．まず，重複レシピ候補群の各レシピをクエリとして，オリジナルレシピ候補群の各レシピに対して，総当たりで文字3-gram集合のJaccard係数を算出する．なお，2つの文字3-gram集合，A，B間のJaccard係数の算出式は5.1式で定義される．

$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{5.1}$$

そして，文字3-gram集合のJaccard係数に基づき重複レシピペア候補を抽出する．なお，実験では材料相違数によるフィルタリングを行った上で重複レシピペア候補を抽出する手法との比較を行う．

$tf \cdot idf$ ベクトルを手がかりとして重複レシピを検出する手法（$tf \cdot idf$）： 本比較手法では，$tf \cdot idf$ ベクトル間のコサイン類似度に基づき，重複レシピペア候補を抽出する．$tf \cdot idf$ は，文書中の単語の出現頻度を示す tf（Term Frequency）と全文書における単語の出現頻度を示す idf（Inverse Document Frequency）を掛け合わせた値である．ここで，tf は5.2式で，idf は5.3式で定義される．

$$tf(t,d) = \frac{n_{t,d}}{\sum_{s \in d} n_{s,d}} \tag{5.2}$$

$$idf(t) = log\frac{N}{df(t)} + 1 \tag{5.3}$$

ただし，5.2式中の $n_{t,d}$ は文書 d 中の単語 t の文書中の出現回数を指し，$\sum_{s \in d} n_{s,d}$ は文書 d 中の単語全ての出現回数を足した値，すなわち文書 d 中の総単語数を指す．また，5.3式中の N は総文書数，$df(t)$ は単語 t が出現する文書数を指す．

まず，実験データ中の全レシピの $tf \cdot idf$ を算出する．そして，重複レシピ候補群の各レシピをクエリとして，オリジナルレシピ候補群の各レシピとの間の $tf \cdot idf$ ベクトルのコサイン類似度を総当たりで算出する．ここで，ベクトル $\vec{d_1}$，ベクトル $\vec{d_2}$ 間のコサイン類似度は5.4式で定義される．

$$cos(\vec{d_1}, \vec{d_2}) = \frac{\sum_{i=1}^{|V|} d_{1i} d_{2i}}{\sqrt{\sum_{i=1}^{|V|} {d_{1i}}^2} \cdot \sqrt{\sum_{i=1}^{|V|} {d_{2i}}^2}} \tag{5.4}$$

実験では，$tf \cdot idf$ ベクトル間のコサイン類似度に基づき重複レシピペア候補を抽出する．なお，実験では材料相違数によるフィルタリングを行った上で重複レシピペア候補を抽出する手法との比較を行う．

表 5.1: 各手法における重複レシピ検出数

材料相違数による フィルタリングあり										材料相違数による フィルタリングなし	
手法	50	100	200	300	手法	50	100	200	300	手法	300
nd_BERT 1 層	32	37	39	43	nd_BERT-WMD 1 層	30	35	43	45	nd_BERT 1 層	1
nd_BERT 2 層	30	33	41	45	nd_BERT-WMD 2 層	30	31	41	**46**	nd_BERT 2 層	1
nd_BERT 3 層	26	32	42	44	nd_BERT-WMD 3 層	25	31	43	**47**	nd_BERT 3 層	1
nd_BERT 4 層	25	34	42	**48**	nd_BERT-WMD 4 層	25	31	43	**48**	nd_BERT 4 層	1
nd_BERT 5 層	26	34	40	**47**	nd_BERT-WMD 5 層	25	32	41	**48**	nd_BERT 5 層	1
nd_BERT 6 層	26	38	41	**46**	nd_BERT-WMD 6 層	28	33	44	**46**	nd_BERT 6 層	1
nd_BERT 7 層	28	35	40	**47**	nd_BERT-WMD 7 層	28	31	45	**46**	nd_BERT 7 層	1
nd_BERT 8 層	28	33	36	44	nd_BERT-WMD 8 層	28	30	41	**46**	nd_BERT 8 層	1
nd_BERT 9 層	27	32	38	**46**	nd_BERT-WMD 9 層	28	31	40	**47**	nd_BERT 9 層	1
nd_BERT 10 層	25	31	40	44	nd_BERT-WMD 10 層	26	31	41	**46**	nd_BERT 10 層	1
nd_BERT 11 層	25	31	38	43	nd_BERT-WMD 11 層	25	30	40	45	nd_BERT 11 層	1
nd_BERT 12 層	25	28	39	42	nd_BERT-WMD 12 層	25	31	41	45	nd_BERT 12 層	1
WMD	26	29	37	42						WMD	1
3-gram	26	34	40	42						3-gram	1
$tf \cdot idf$	25	32	42	42						$tf \cdot idf$	1

5.5 実験結果

表 5.1 に提案手法における調理手順テキスト間の距離上位 50 件，100 件，200 件，300 件における重複レシピ検出数を示す．

ただし，材料相違数による重複レシピペア候補のフィルタリングを行わない手法については，上位 300 件における重複レシピ検出数が提案手法のいずれの層よりも低くなっているため，上位 300 件における重複レシピ検出数のみを示す．

材料相違数によるフィルタリングの有効性の検証: 材料相違数によるフィルタリングの有効性を検出するため，nd_BERT の 1 層〜12 層，WMD，3-gram，$tf \cdot idf$ 各手法の上位 300 件における重複レシピ検出数を基に有効性を検証した．その結果，重複レシピペア候補を材料相違数を用いてフィルタリングする手法が有効であることが示された（フィルタリングを行う場合とフィルタリングを行わない場合の間で t-検定（両側検定，有意水準 1%，p=3.803 $\times 10^{-20}$）で有意に向上）．材料相違数によるフィルタリングを行わない場合，材料相違数によるフィルタリングを行う手法に比べて，大幅に重複レシピの検出数が減少した．これは材料相違数によるフィルタリングを行わない場合，調理手順テキストが完全に，もしくは大部分が一致しているものの，使用されている材料が異なるレシピが抽出されたためである．

提案手法の有効性の検証: 表 5.1 の上位 300 件における重複レシピ検出結果に着目すると，nd_BERT および nd_BERT-WMD において，最も多い層で 48 件の重複レシピが検出された．それ以外の層においても，WMD，3-gram，$tf \cdot idf$ と同等もしくはそれ以上の検出数となっている．この結果より，提案手法である nd_BERT および nd_BERT-WMD の有効性が示された．とりわけ，比較手法に用いた 3-gram は過去の重複レシピ検出に関する研究において，最も良い結果を残しているが，提案手法である nd_BERT および nd_BERT-WMD は，それを上回る結果となった．

WMD によるリランキングの有効性の検証: 続いて，WMD によるリランキングの有効性を検証するため，nd_BERT-WMD の 1 層〜12 層と nd_BERT の 1 層〜12 層の上位 300 件における重複レシピ検出数を基に有効性を検証した結果，有意差（nd_BERT-WMD と nd_BERT の間で t-検定（両側検定，有意水準 1%，p=0.003）で有意に向上）が確認でき，WMD によるリランキングの有効性が示された．

上記の結果から，材料相違数によるフィルタリングの有効性，提案手法である nd_BERT および nd_BERT-WMD の有効性，WMD によるリランキングを行う手法の有効性が示された．

ここで，本実験で検出された 55 件の重複レシピを 1 とすると，重複レシピの検出結果が最も優れていた nd_BERT-WMD 4 層における上位 50 件，100 件，200 件，300 件における再現率は，上位 50 件が 0.455，上位 100 件が 0.564，上位 200 件が 0.782，上位 300 件が 0.873 となった．また，精度は上位 50 件が 0.500，上位 100 件が 0.310，上位 200 件が 0.215，上位 300 件が 0.160 となった．この結果より，単に効率よく重複レシピを見つけるという点においては，上位 50 件までを評価対象とするのみで良い．一方で重複レシピは投稿型レシピサイトのユーザビリティを損なうリスクを孕んでいるため，可能な限り多くの重複レシピを検出し，削除することが望ましい．重複レシピか否かを判断するのに要する時間は 1 分程度である．すなわち，300 件に対してアノテーションを行った場合の所要時間は，5 時間程度である．上述した通り，400 件，500 件とアノテーションの件数を増加させても，重複レシピの数は大きく変化しない．以上の議論を踏まえ，重複レシピの検出カバレッジおよびレシピペアに対して人手でアノテーションを行う労力のバランスを考えると，上位 300 件までを評価対象とするのが妥当といえる．

5.6 考察

本節では，本研究の考察として，5.6.1 項において，WMD を用いてリランキングを行う nd_BERT-WMD の有効性および課題について検証した結果，5.6.2 項において，重複レシピの検出に有効な BERT の層についての分析結果をそれぞれ述べる．

5.6.1 nd_BERT-WMD の有効性および課題の検証

nd_BERT-WMD の有効性　表 5.2 に提案手法を用いることで重複レシピの検出結果が改善したレシピペアの調理手順テキストの例を示す．表 5.2 のレシピペアは，nd_BERT では 7 つの層でしか検出できなかったのに対して，nd_BERT-WMD では全ての層で検出できた．

表 5.2: nd_BERT-WMD で重複レシピの検出結果が改善したレシピペアの例

	調理手順テキスト
重複レシピ	鍋にざく切りにしたじゃがいも、人参、大根、えのき、昆布、水１５０ｃｃを入れて火にかける　煮たったら煮干し粉を入れて具に火が通るまで加熱する　火を止めて味噌をとく
オリジナルレシピ	鍋にざく切りにしたじゃがいも、大根、人参、えのき、水１５０ｃｃを入れて火にかける　煮たったら煮干し粉を入れて具に火が通るまで加熱する　火を止めて味噌をとく

調理手順テキストに着目すると，重複レシピに昆布が追加されており，重複レシピとオリジナルレシピの間で人参と大根が言及される順序が異なっているだけの違いである．nd_BERT において検出を行えなかった 5 つの層における調理手順テキスト間の距離の順位に着目すると，それぞれ 311 位，527 位，556 位，452 位，338 位となっている．これらが WMD によるリランキングを行ったことで，nd_BERT-WMD では，154 位，182 位，189 位，170 位，144 位でそれぞれ検出された．WMD は文書中の単語同士を対応付けることにより，文書間の距離を算出する手法である．そのため，調理手順テキストにおける単語の順序の入れ換えに頑健である．しかし，BERT では位置埋め込みを用いるため，同じ単語であっても文中の出現位置によって抽出される埋め込み表現が異なる．表 5.2 に挙げた例では，人参と大根の記述されている順序が異なり，昆布が新たに追加されている．このことが影響して，WMD でリランキングを行う nd_BERT-WMD において検出結果が改善したと考えられる．

nd_BERT-WMD の課題　続いて，表 5.3 に比較手法では検出できたが，nd_BERT および nd_BERT-WMD において検出結果が悪化したレシピペアの調理手順テキストの例を示す．表 5.3 に示したレシピペアは，nd_BERT では 8 つの層で検出できたのに対して，nd_BERT-WMD では上位 300 位以内で検出できなかった．

表 5.3: nd_BERT-WMD で重複レシピの検出結果が悪化したレシピペアの例

	調理手順テキスト
重複レシピ	かぼちゃをレンジで５分くらい加熱し、マッシュする　フライパンで細切りにしたベーコン、みじん切りにしたたまねぎを炒める　１と２、塩こしょう、粉チーズを混ぜ合わせ、小判形に丸める　小麦粉、卵、パン粉の順に衣をつけ、１８０度の油できつね色になるまで揚げる
オリジナルレシピ	かぼちゃをざく切りにし、レンジで２～３分くらい加熱し、マッシュする　フライパンでみじん切りにしたベーコン、たまねぎを炒める　１と２、チーズ、塩こしょうを混ぜ合わせ、小判形にする　小麦粉、溶き卵、パン粉の順に衣をつけ、油できつね色になるまで揚げる

調理手順テキストに着目すると，重複レシピでは「レンジで 5 分ぐらい加熱し」と記述されていた部分が，オリジナルレシピでは「2〜3 分」に変更されている．また，ベーコンの調理方法が重複レシピと非重複レシピで異なっており，チーズが粉チーズ，卵が溶き卵に入れ替わっている等，様々な部分で調理手順テキストに変化がある．このレシピペアでは,「ざく切り」や「細切り」,「180 度」のように一方のレシピにのみ出現する単語が多い．WMD では，単語と単語を紐付けすることにより，文書間の距離を求める．そのため，一方のレシピの単語と意味的な距離が遠い単語を追加された場合に WMD が非常に大きく算出される．このことにより，WMD を用いてリランキングを行う提案手法において検出結果が悪化したと考えられる．

5.6.2　重複レシピの検出に有効な BERT の層の分析

表 5.1 の上位 300 件における重複レシピの検出数に着目すると，nd_BERT では 1 層から 3 層の浅い層と 10 層から 12 層までの深い層にかけて，重複レシピの検出数が他の層と比べて落ち込んでいるのに対して，調理手順テキスト間の WMD に基づいてリランキングを行う nd_BERT-WMD ではどの層においても，重複レシピの検出数が安定し

た結果となっている．ただし，4 層から 9 層にかけての中間層においては，nd_BERT と nd_BERT-WMD の重複レシピの検出件数が同等になっている．一般に言語処理における深層学習モデルでは，浅い層では形態的な特徴が，深い層では意味的な特徴が抽出されることが知られている [9, 23]．すなわち，浅い層では形態的に類似したレシピペアが，深い層では意味的に類似したレシピペアが上位にランキングされる．これを踏まえると，4 層から 9 層にかけての中間層では，形態的な特徴と意味的な特徴の双方を扱えていることにより，WMD によるリランキングを行わない場合でも，重複レシピの検出結果が優れていたといえる．

また，nd_BERT-WMD では，調理手順テキスト間の WMD に基づくリランキングを行うことにより，浅い層においては欠けていた意味的な特徴が，深い層においては欠けていた形態的な特徴がそれぞれ補完され，どの層においても安定して重複レシピの検出を行えたと考えられる．以上の議論から，重複レシピの検出タスクにおいては，形態的な特徴と意味的な特徴の双方を扱える層，つまり 4 層から 9 層にかけての中間層が適しているといえる．

第6章　おわりに

6.1　本研究のまとめ

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．例えば，ユーザがレシピサイトを利用してレシピを検索した際，検索結果に調理手順テキストが重複しているレシピが複数表示され，かつその中にユーザが求めているレシピがない場合，ユーザが目的のレシピを選択する過程において，余分な時間的コストを掛けることになる．また，こうした事象が繰り返し発生し，ユーザがレシピサイトを利用しなくなった場合，レシピサイトの運営者にとっても不利益を被る結果に繋がる．そこで本研究では，重複したレシピ，すなわち重複レシピの検出手法を提案し，上記の問題を解決する上での一助とすることを目的とする．重複レシピは，レシピ投稿者が他者のレシピを剽窃することによって作成されるものである．こうした剽窃に関する問題は，投稿型レシピサイトのみならず，学術論文や小説，楽曲など，幅広い分野が抱えている．また，近年ではソーシャル・ネットワーキング・サービス（SNS）上でも，他人の投稿を剽窃した投稿が問題視されるケースもある．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．以上の研究背景と目的を 1 章において述べた．

2 章では，重複レシピの検出に関する研究，文書の剽窃検出に関する研究，文書の埋め込み表現の抽出に関する研究，単語の埋め込み表現を用いて文書間の類似性を算出した研究について論じ，本研究との相違点を明らかにするとともに，本研究の位置づけを示した．

3 章では，本研究の提案手法である 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手法について述べた．提案手法では，以下の 3 点をねらいとした．

- 文書の意味的な等価性を判定するタスクである Microsoft Research Paraphrase Corpus（MRPC）[5] をはじめとする，複数の自然言語処理のタスクで最高精度を達成した BERT を用いて調理手順テキストの埋め込み表現を抽出することで，調理手

順テキストの意味についても捉えた重複レシピの検出を行う

- 材料相違数によるフィルタリングを行うことで，調理手順テキストは共通しているものの，材料がまったく異なる料理について記述したレシピペアを重複レシピとして誤検出することを防ぐ

- WMD に基づいて，重複レシピペア候補のリランキングを行うことで，BERT モデルを用いて調理手順テキストの埋め込み表現を抽出した場合に調理手順テキスト中の単語間の意味的な対応が考慮されない点を補う

提案手法では，調理手順テキストの埋め込み表現間の距離に基づく重複レシピペア候補のランキング，材料相違数に基づく重複レシピペア候補のフィルタリング，WMD に基づく重複レシピペア候補のリランキングを行うことにより重複レシピの検出を行う．

4 章では，実験に用いるデータセットおよび重複レシピペア候補のアノテーション基準について述べた．実験では，楽天株式会社より提供を受けた楽天レシピのデータセットを用いた．アノテーション基準を用いて行った事前実験の結果について述べ，3 名の実験参加者を雇用し，著者を含めた 4 名のアノテータでアノテーションを行った結果を示した．

5 章では，提案手法である 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検出手法の有効性の検証結果および重複レシピの検出における材料相違数によるフィルタリングの有効性の検証を行った結果を示した．評価の結果，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 によるフィルタリングを行う手法と行わない手法における有意水準 1%の有意差が確認でき，提案手法の有効性を示した．

6.2 今後の課題

今後の課題として，調理に直接関係のない文を除去することにより，調理手順テキストから抽出される埋め込み表現から重複レシピの検出に影響を及ぼす要素を削除することが挙げられる．本研究の考察で示したとおり，BERT では位置埋め込みを用いるため，同じ単語であっても文中の出現位置によって抽出される埋め込み表現が異なる．そのため，記述の順序が入れ替えられたり,「できあがり」や「完成です」などの文言が新たに加えられた場合，BERT を用いて抽出される調理手順テキストの埋め込み表現に違いが出る．

また，重複レシピの検出に有効なBERTの層の分析結果を踏まえ，複数の層を組み合わせて重複レシピの検出を行うことが挙げられる．考察において示したとおり，言語処理における深層学習モデルでは，浅い層では形態的な特徴が，深い層では意味的な特徴が抽出されることが知られている．これらの特徴を生かして，1層と12層など，異なる層から抽出される調理手順テキストの埋め込み表現を用いることにより，重複レシピの検出結果の改善が期待される．

本研究では一般に公開されているBERTの事前学習モデルを用いた．今回用いた事前学習モデルは日本語Wikipediaを用いて学習されている．今後，BERTモデルをレシピデータによって学習することにより重複レシピ検出精度の向上を図りたいと考えている．

謝辞

はじめに，本研究を進めるにあたって，時に厳しく時に優しく，私と向き合って研究指導を頂いた筑波大学図書館情報メディア系関洋平准教授に深く感謝の意を表します．関准教授には，筑波大学学群学生であったときから3年間ご指導を頂きました．研究面だけではなく，人間的にも大きく成長できたのは，関准教授の指導の賜物です．この場をお借りして深く御礼を申し上げます．

副指導教員を務めて頂いた筑波大学図書館情報メディア系高久雅生准教授には，関准教授がサバティカルで不在の1年間主指導教員として指導して頂きました．特に博士前期課程1年次では，思ったような研究成果が出ず，思い悩む時期が長かったですが，高久准教授の何者をも包み込む優しい笑顔に励まされ，折れずに研究を続けることができました．この場をお借りして深く御礼を申し上げます．

楽天株式会社楽天技術研究所 平手勇宇様には，論文執筆や研究の進め方について，目から鱗が落ちるような助言を頂きました．この研究を行うことができたのも，平手様のお陰です．また，楽天株式会社にはインターンシップの受け入れもして頂きました．この場をお借りして感謝の言葉を申し上げます．

京都大学情報学研究科の杉山一成特定准教授には，しばしばゼミにご参加頂き，研究を進めていく上での重要なアドバイスを頂きました．この場をお借りして感謝の言葉を申し上げます．

また，コミュニケーション理解研究室の同期である趙康康さんとは，互いに励まし合いながら2年間研究を続けて来ました．趙さんのいつの日も笑顔で前向きな性格に励まされることも多々ありました．本当にありがとうございました．

同じくコミュニケーション理解研究室の劉依泓さん，上田悠登さん，小久保千裕さん，齊藤幸乃さん，冨平準喜さん，石田哲也さん，谷口愛依さん，中山聖司さんには，先輩として至らない私を多岐に渡り支えて頂きました．心より感謝の言葉を申し上げます．

学部時代にコミュニケーション理解研究室の同期であった安藤有生さん，南澤亜樹さんには，修士論文の執筆に向けて激励の言葉を頂きました．お2人のお言葉のお陰で本論文の執筆を完遂することができました．本当にありがとうございました．

同じ 7D 棟 140 号室で苦楽を共にしたコンテンツ工学研究室の皆様には，多くの助言を頂きました．特に同期の稲福和史さんとは，切磋琢磨をしあい研究を進めることができました．ここに感謝いたします．

本研究の一部は，科学研究費補助金基盤研究 B（課題番号 19H04420）の助成を受けて遂行されました．また，本研究で用いたレシピデータは楽天株式会社 楽天技術研究所からご提供頂いたものです．ここに深く感謝の意を示します．

参考文献

[1] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France, 2017. OpenReview.net.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146, 2017.

[3] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, Vol. 20, No. 1, p. 37, 1960.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[5] William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing*, pp. 9–16, Jeju Island, Korea, 2005. Association for Computational Linguistics.

[6] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, Vol. 76, No. 5, pp. 378–382, 1971.

[7] Vivek Gupta, Harish Karnick, Ashendra Bansal, and Pradhuman Jhala. Product classification in e-commerce using distributional semantics. In *Proceedings of the 26th International Conference on Computational Linguistics*, pp. 536–546, Osaka, Japan, 2016. The COLING 2016 Organizing Committee.

[8] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[9] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3651–3657, Florence, Italy, 2019. Association for Computational Linguistics.

[10] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017.

[11] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, pp. 3294–3302, Montreal, Canada, 2015. MIT Press.

[12] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, Brussels, Belgium, 2018. Association for Computational Linguistics.

[13] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, pp. 957–966, Lille, France, 2015. Association for Computing Machinery.

[14] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, Vol. 33, No. 1, pp. 159–174, 1977.

[15] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*, pp. 1188–1196, Beijing, China, 2014. Association for Computing Machinery.

[16] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Learning context-sensitive word embeddings with neural tensor skip-gram model. In *Proceedings of the 24th In-*

ternational Conference on Artificial Intelligence, pp. 1284–1290, Buenos Aires, Argentina, 2015. Association for Computing Machinery.

[17] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, BC, Canada, 2018. OpenReview.net.

[18] Dheeraj Mekala, Vivek Gupta, Bhargavi Paranjape, and Harish Karnick. SCDV : Sparse composite document vectors using soft clustering over distributional representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 659–669, Copenhagen, Denmark, 2017. Association for Computational Linguistics.

[19] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pp. 3111–3119, Lake Tahoe, Nevada, USA, 2013. Curran Associates Inc.

[20] Masaki Oguni, Yohei Seki, and Yu Hirate. Character 3-gram mover's distance: An effective method for detecting near-duplicate japanese-language recipes. *arXiv preprint arXiv:1912.05171*, 2019.

[21] Masaki Oguni, Yohei Seki, Risako Shimada, and Yu Hirate. Method for detecting near-duplicate recipe creators based on cooking instructions and food images. In *Proceedings of the 9th Workshop on Multimedia for Cooking and Eating Activities in conjunction with the 2017 International Joint Conference on Artificial Intelligence*, pp. 49–54, Melbourne, Australia, 2017. Association for Computing Machinery.

[22] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

[23] Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1499–1509, Brussels, Belgium, 2018. Association for Computational Linguistics.

[24] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, Vol. 40, No. 2, pp. 99–121, 2000.

[25] M. Schuster and K.K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, Vol. 45, No. 11, pp. 2673–2681, 1997.

[26] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany, 2016. Association for Computational Linguistics.

[27] Kohei Sugawara, Hayato Kobayashi, and Masajiro Iwasaki. On approximately searching for similar word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2265–2275, Berlin, Germany, 2016. Association for Computational Linguistics.

[28] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, pp. 3104–3112, Montreal, Canada, 2014. MIT Press.

[29] Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. Juman++: A morphological analysis toolkit for scriptio continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 54–59, Brussels, Belgium, 2018. Association for Computational Linguistics.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In

Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010, Long Beach, California, USA, 2017. Curran Associates Inc.

[31] Xiaojie Wang, Zhicheng Dou, Tetsuya Sakai, and Ji-Rong Wen. Evaluating search result diversity using intent hierarchies. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 415–424, Pisa, Italy, 2016. Association for Computing Machinery.

[32] 高橋勇, 宮川勝年, 小高知宏, 白井治彦, 黒岩丈介, 小倉久和. Web サイトからの剽窃レポート発見支援システム. 電子情報通信学会論文誌. D, 情報・システム, Vol. 90, No. 11, pp. 2989–2999, 2007.

[33] 小邦将輝, Nio Lasguido, 平手勇宇, 関洋平. 調理手順テキストと料理画像の特徴量の最近傍探索に基づく重複レシピの検出手法. 電子情報通信学会技術報告, 第 118 巻, pp. 19–24, 茨城，つくば, 2018. 電子情報通信学会.

[34] 小邦将輝, 関洋平, 平手勇宇. 重複レシピの検出における単語の分散表現と文字 n-gram の分散表現の比較. ARG 第 14 回 Web インテリジェンスとインタラクション研究会, pp. 29–32, 兵庫，神戸, 2019. ARG Web インテリジェンスとインタラクション研究会.

[35] 小邦将輝, 島田理紗子, 平手勇宇, 杉山一成, 関洋平. レシピの素性を用いた重複レシピ判別の検証. 第 10 回データ工学と情報マネジメントに関するフォーラム, DEIM2018-J2-2, 福井，あわら, 2018. 日本データベース学会.

[36] 小高知宏, 村田哲也, 高建斌, 諏訪いずみ, 白井治彦, 高橋勇, 黒岩丈介, 小倉久和. n-gram を用いた学生レポート評価手法の提案. 電子情報通信学会論文誌. D-I, 情報・システム, I-情報処理, Vol. 86, No. 9, pp. 702–705, 2003.

[37] 柴田知秀, 河原大輔, 黒橋禎夫. Bert による日本語構文解析の精度向上. 言語処理学会第 25 回年次大会発表論文集, pp. 205–208, 愛知，名古屋, 2019. 言語処理学会.

[38] 島田理紗子, 小邦将輝, 平手勇宇, 関洋平. 重複する料理レシピを判別するためのコーパスの構築. ARG 第 11 回 Web インテリジェンスとインタラクション研究会, pp. 47–50, 東京, 2017. ARG Web インテリジェンスとインタラクション研究会.

[39] 久保遥, 関洋平. 投稿型レシピサイトを横断した重複レシピの判別. 第 8 回データ工学と情報マネジメントに関するフォーラム, DEIM2016-C8-3, 福岡, 2016. 日本データベース学会.

発表論文

査読付国際会議論文

(1) Masaki Oguni, Yohei Seki, Risako Shimada, and Yu Hirate. Method for detecting near-duplicate recipe creators based on cooking instructions and food images. In *Proceedings of the 9th Workshop on Multimedia for Cooking and Eating Activities in conjunction with the 2017 International Joint Conference on Artificial Intelligence*, pp. 49-54, Melbourne, Australia, 2017. Association for Computing Machinery.

国内会議論文（和文）

(1) 小邦将輝，関洋平，平手勇宇．重複レシピの検出における単語の分散表現と文字 N-gram の分散表現の比較．ARG 第 14 回 Web インテリジェンスとインタラクション研究会，pp. 29–32，兵庫，神戸，2019，ARG Web インテリジェンスとインタラクション研究会．(**学生奨励賞，学生優秀ポスター発表賞 受賞**)

(2) 小邦将輝，Lasguido Nio，平手勇宇，関洋平．調理手順テキストと料理画像の特徴量の最近傍探索に基づく重複レシピの検出手法．調理手順テキストと料理画像の特徴量の最近傍探索に基づく重複レシピの検出手法．電子情報通信学会技術報告，Vol. 118，pp. 19–24，茨城，つくば，2018，電子情報通信学会．

(3) 島田理紗子，小邦将輝，平手勇宇，関洋平．重複する料理レシピを判別するためのコーパスの構築．Web インテリジェンスとインタラクション研究会第 6 回ステージ発表．(**第 6 回ステージ発表採択 採択率 21.7%**)

(4) 小久保千裕，小邦将輝，関洋平．地域特有の埋め込み表現を用いたイベント参加地域の推定．ARG 第 15 回 Web インテリジェンスとインタラクション研究会，pp. 25–28，神奈川，横浜，2019，ARG Web インテリジェンスとインタラクション研究会．

(5) 小久保千裕，小邦将輝，関洋平．イベント参加地域推定のための単語埋め込み表現の拡張．第 12 回データ工学と情報マネジメントに関するフォーラム，4p, 福島，磐梯熱海，2020，日本データベース学会．

国内会議論文（英文）

(1) Kangkang Zhao, Masaki Oguni, Yohei Seki, and Kazunari Sugiyama. A method for classifying temporal relations using attention-based neural networks. IEICE Technical Report, Vol. 119, No. 212, pp. 93-98, Tokyo, Japan, 2019.

投稿中の査読付論文

査読付学術雑誌論文

(1) 小邦将輝，関洋平，平手勇宇．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．情報処理学会論文誌データベース (TOD), Vol. 13, No. 2, 13p, 2020.

査読付国際会議論文

(1) Kangkang Zhao, Yohei Seki, Masaki Oguni, and Kazunari Sugiyama. A method for classifying temporal relations with attention to the context of questioned time-Event entities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5p, Seattle, USA, 2020. Association for Computational Linguistics.

査読なし英語論文

(1) Masaki Oguni, Yohei Seki, Yu Hirate. Character 3-gram Mover's Distance: An effective method for detecting near-duplicate japanese-language recipes. *arXiv preprint arXiv:1912.05171*, 5p, 2019.

社会貢献活動

(1) 水戸市「今日から始める Twitter 講座 基本編 講師」. 2018 年 9 月.（茨城新聞に掲載：澤田将生．ツイッターで顧客獲得 水戸市が講座 経営者ら注意点学ぶ．茨城新聞．2018 年 9 月 18 日，朝刊.）

(2) 水戸市「今日から始める Twitter 講座 応用編 講師」. 2018 年 10 月.