ARG WI2 No.14, 2019
WI2-2019-07

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

小邦 将輝[†,a]　関 洋平[††,b]　平手 勇宇[†††,c]

† 筑波大学大学院図書館情報メディア研究科　†† 筑波大学図書館情報メディア系
††† 楽天株式会社楽天技術研究所

a) *s1821613@s.tsukuba.ac.jp*　b) *yohei@slis.tsukuba.ac.jp*　c) *yu.hirate@rakuten.com*

概要　投稿型レシピサイトには，調理手順テキストなどの料理レシピの構成要素が他のレシピと同一のレシピ（重複レシピ）が存在する．本研究では，単語の分散表現間の距離に基づいて文書間の距離を算出する手法である Word Mover's Distance を文字 3-gram の分散表現へと応用した手法を提案する．評価実験では，約 121 万件のレシピから単語の分散表現と文字 3-gram の分散表現を Skip-gram Model with Negative Sampling，fastText の 2 手法を用いて学習し，重複レシピペア候補を抽出する．そして，重複レシピペア候補へのアノテーションを行い，重複レシピ検出手法の評価を行う．実験の結果，単語の分散表現を用いた際には検出できなかった重複レシピが，文字 3-gram の分散表現を用いることによって検出できることを確認した．

キーワード　重複レシピ，重複検出，Word Mover's Distance，文字 N-gram

1　はじめに

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，調理手順テキストなどが他のレシピと完全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，これらのレシピを「**重複レシピ**」，重複レシピが剽窃する元となったと考えられるレシピを「**オリジナルレシピ**」とする．また，重複レシピとオリジナルレシピのペアについて「**重複レシピペア**」とする．

Kusner et al.[1] が提案した単語の分散表現間の距離をもとに文書間の距離を求める Word Mover's Distance (WMD) は，類似文書検索において高い精度を残した．また小高ら [2] は日本語の重複検出には文字 3-gram が有効だとする知見を示した．以上より，本研究では WMD を文字 3-gram の分散表現へと応用した手法を提案する．

本論文では，単語および文字 3-gram の分散表現を用いて重複レシピを検出し，双方の手法を比較した結果を示す．また，分散表現を Skip-gram Model with Negative Sampling (SGNS) [3]，fastText[4] の 2 手法で学習し，分散表現の学習手法による結果の違いを分析する．

2　関連研究

著者らの先行研究 [5] では，小高ら [2] が示した日本語の重複検出に関する知見に基づき，重複レシピの検出に文字 3-gram を用いた．しかし，レシピ中には言い換えが可能な材料や，複数の表記方法を持つ材料が存在する．またユーザによって作成されたレシピには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，重複レシピを検出できない．そこで，文字 3-gram の分散表現を用いて，レシピ中の誤字や脱字に対して頑健な重複レシピの検出手法を提案する．

文書の重複検出に関する研究として，著者らの先行研究 [6] では Mekala et al.[7] が提案した Sparse Composite Document Vectors に基づいた手法を提案した．この手法では，文書の特徴量を抽出し，特徴量間の距離に基づき文書間の類似度を求めた．重複レシピの検出は，ユーザが投稿したレシピに対して，重複か否かを判別し，サイト掲載の可否を決定する特性上，判別の根拠を明示することが望ましい．そこで，Kusner et al.[1] が提案した単語の分散表現間の距離に基づいて直感的に文書間の距離を算出可能な WMD を用いる．

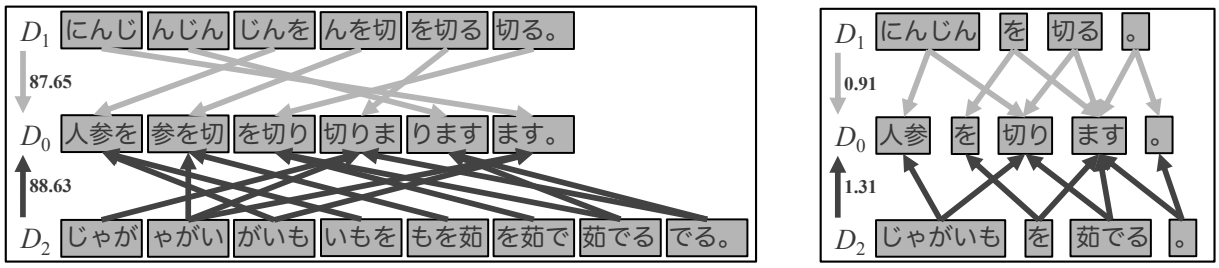
3　提案手法

本研究では，WMD を文字 3-gram の分散表現へと応用した手法を提案する．図 1 (a) に提案手法における調理手順テキスト間の距離の算出手法を示す．Kusner et al. が提案した手法 (図 1(b)) では，単語同士の置き換えコストの総和を文書間の距離と定義した．なお，単語同士の置き換えコストとは，単語の分散表現間のユークリッド距離のことを指す．

提案手法では，WMD を文字 3-gram の分散表現へと応用し，調理手順テキスト間の距離を文字 3-gram 間の距離の総和とする．これにより言い換えや書き換え，誤字・脱字が含まれる場合においても頑健に重複レシピを検出できると考えられる．

4　重複レシピの検出における単語の分散表現と文字 N-gram の分散表現の比較実験

本章では提案手法の有効性の検証を目的とした実験について示す．実験では，提案手法と 4.2.2 項で示す比較



(a) 提案手法　　**(b)** 比較手法

図 1: 提案手法と比較手法における調理手順テキスト間の距離の算出方法

手法を用いて，重複レシピ検出精度の比較を行う．

4.1 データセット

本実験には，楽天レシピのデータセットを用いる．データセットは，2010 年 6 月 30 日から 2016 年 11 月 8 日までの間に投稿された 1,214,650 件のレシピについて，調理手順，材料などの情報から構成されている．

実験では，文字 3-gram の分散表現を学習する際のデータとして，2010 年 6 月 30 日から 2016 年 10 月 31 日までに投稿された 1,210,612 件のレシピを用いる．またテストデータとして，2016 年 11 月 1 日から 2016 年 11 月 8 日の間に投稿された 4,038 件のレシピを用いる．

4.2 実験方法

4.2.1 実験手順

実験では，訓練データを用いて，単語および文字 3-gram の分散表現を SGNS，fastText の 2 手法で学習する[1]．なお，単語の分散表現については比較手法で用いるために学習を行った．続いて，学習した 4 つの分散表現を用い，テストデータ中の各レシピをクエリとして，訓練データ中から調理手順テキスト間の距離上位 10 件のレシピを抽出する．そして，材料相違数（4.2.3 項）に基づき重複レシピペア候補の中からアノテーション対象レシピペアを抽出する．この結果，アノテーション対象レシピは，異なり 1,983 件のレシピペアとなった．最後に対象となるレシピペアについて，アノテーション（4.2.4 項）を行い，重複レシピ検出手法の評価を行う．

4.2.2 比較手法

本実験では，比較手法として Kusner et al. によって提案された WMD を用いる (図 1(b))．実験では，本手法を比較手法として用いることにより，提案手法および比較手法の有効性について検証する．

4.2.3 材料相違数の算出方法

本項では，材料相違数の算出手法を述べる．レシピ中で用いられる材料の中には，同じ材料であるにもかかわらず，言い換えが可能な材料や，複数の表記方法を持つ材料が存在する．そこで，著者らの先行研究 [6] に基づき，以下の手順で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の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，括弧内の文字は削除する．
(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．
(3) 重複レシピ候補の各材料について，単語の分散表現をもとに，類似単語を検索する．このとき類似単語の検索結果上位 3 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，両レシピの材料リストから削除する．
(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．

著者らの先行研究では，材料相違数が 0 の場合に重複レシピとして検出した．しかし，上記のアルゴリズムでは，材料に「カットしたわかめ」，もう一方に「カットわかめ」などと記載されている場合，「カットしたわかめ」が 1 つの材料として認識されないため，正しく類似材料の検索を行えない．そのため，材料相違数が 2 となる．すなわち，先行研究の基準では，重複レシピとして検出できない．そこで，検出可能な重複レシピの範囲を広げることを目的として，材料相違数が 2 以下のレシピペアをアノテーション対象とする．

4.2.4 レシピのアノテーション基準

本研究では，アノテーションを行う際の基準に，先行研究 [6] と同様のものを用いる．以下にアノテーション基準の一例を示す．アノテーションを行う際には，以下の 4 段階の基準に基づき，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する．

- 重複
  - 材料が完全に一致しており，調理手順テキストも文末表現等を除き一致しているもの
- 非重複 A
  - 同じ料理で，材料についても共通している部分があり，調理手順に共通点が見られるもの

[1] 学習時のパラメータ: 分散表現の次元数は 100 次元とし，窓幅を 15 に設定した．最低頻度について，単語の分散表現については 1，文字 3-gram の分散表現については 10 とした．

表 1: レシピのアノテーション結果

手法		重複	非重複 A	非重複 B	非重複 C	合計
SGNS	文字 3-gram	46 (4.17%)	424 (38.41%)	331 (29.98%)	303 (27.45%)	1,104
	word	**47** (3.49%)	470 (34.89%)	382 (28.36%)	448 (33.26%)	1,347
fastText	文字 3-gram	**47** (4.38%)	414 (38.55%)	301 (28.03%)	312 (29.05%)	1,074
	word	46 (4.38%)	410 (39.05%)	281 (26.76%)	313 (29.81%)	1,050
異なり数		50 (2.52%)	575 (29.00%)	535 (26.98%)	823 (41.50%)	1,983

- 非重複 B
  - 主要な材料が異なっているが，異なる材料についての処理を除けば，調理手順に共通点が見られるもの
- 非重複 C
  - 異なる料理で，調理手順，材料についても異なるもの

なお，本基準を用いて先行研究でアノテーションを行った結果，本論文の第一著者と実験参加者[2]の回答の一致度を示す Cohen の κ 係数 [8] が 0.903 (Almost perfect agreement[9]) となり，十分に信頼のおけるアノテーション基準であることが示されたため，本実験では第一著者がアノテーションを担当した．

4.3 実験結果

表 1 にアノテーション結果を示す．実験の結果，fastText については，文字 3-gram の分散表現を用いたときに，より多くの重複レシピを検出できるのに対して，SGNS については，単語の分散表現を用いたときに重複レシピの検出数が多いことがわかった．また，各ラベルの割合に着目すると，いずれの場合でも非重複 A の割合が最も高くなっていることがわかる．この結果は，調理手順テキストを書き換えず，材料を入れ替えただけの質の低いレシピが多く存在することを示している．

材料相違数に着目すると，材料相違数 0 のとき 33 件，1 のとき 2 件，2 のとき 15 件の重複レシピが確認された．先行研究 [6] では，材料相違数が 0 のとき重複レシピとして検出を行っていたが，材料相違数の閾値を上げることで，さらに多くの重複レシピを検出できることがわかった．また，先行研究で調理手順テキストを基に検出した重複レシピは 28 件であったが，提案手法では材料相違数 0 のときに 33 件の重複レシピを検出できた．

提案手法と比較手法の間で検出できた重複レシピの差異について検証したところ，一方の手法で検出できなかった重複レシピをもう一方の手法を用いた際に検出できていることがわかった．提案手法でのみ検出できた重複レシピペアには，一方のレシピに誤字が含まれている，「もやし」から「1.」など一方のレシピで材料が省略形に置き換わっている，「カットした茎わかめ」と「カットわかめ」など文字列的に見ると類似しているにも関わらず，単語として見たときに距離が遠くなっている[3]という特徴があった．一方，比較手法でのみ検出できた重複レシピペアには，語順の入れ替え，「粗みじん切りにする」を「みじん切りにしておく」のような書き換え，文単位での類似した表現への言い換えが見られた．

実験の結果，文字 3-gram の分散表現を用いることで，単語の分散表現のみを用いるときに比べ，重複レシピの検出範囲を広げられることがわかった．しかし，これらの手法を単に組み合わせるだけでは，重複レシピの検出数は増加する一方で，非重複レシピの誤検出が増加する．検出した重複レシピの差異を分析した結果，提案手法の弱点として，語順の入れ替えや文単位での類似した表現への言い換えが挙げられる．そこで，材料などの語順の入れ替えが可能なものを，記号に差し替えることで，語順の入れ替えにも頑健に対処できると考えられる．

5 考察

本研究における考察として，機械学習手法を用いた調理手順テキスト間の距離の算出精度の評価について述べる．4 章の実験において，SGNS と fastText を分散表現学習手法として用い，単語および文字 3-gram の分散表現を学習した．そして，調理手順テキスト間の距離および材料相違数をもとに重複レシピペア候補を抽出した．

このとき，各重複レシピペア候補の調理手順テキスト間の距離には大きく隔たりがある．すなわち，重複レシピペアの場合調理手順テキスト間の距離が近くなり，非重複レシピペアの場合調理手順テキスト間の距離が遠くなるように調理手順テキスト間の距離を算出できていれば，高精度での重複レシピの検出が可能になると考えられる．そこで，機械学習手法を用いて，調理手順テキスト間の距離および材料相違数を素性として分類を行い，分類精度に基づいて学習した分散表現について評価する．

評価を行う際には，重複レシピペアを正例，非重複レシピペアを負例とし，ロジスティック回帰，サポートベクターマシン（線形カーネル，RBF カーネル），ランダムフォレスト，ナイーブベイズの 5 手法を用いて，調理手順テキスト間の距離および材料相違数を素性として分類を行う．一般に，不均衡データを用いた 2 値分類では，訓練データ数の差が結果に影響を及ぼすことが知られている [10]．そこで，under sampling を行い，非重複レシ

[2] 40 代の女性

[3] 形態素解析を行う際に，「カットわかめ」は 1 単語と認識されるが，「カットした茎わかめ」は，カット，し，た，茎わかめとなる．

表 2: 機械学習手法を用いた重複レシピ分類結果

手法		ロジスティック回帰			SVM (線形)			SVM (RBF)			ランダムフォレスト			ナイーブベイズ		
		F 値	再現率	精度	F 値	再現率	精度	F 値	再現率	精度	F 値	再現率	精度	F 値	再現率	精度
SGNS	文字 3-gram	**0.81**	0.89	0.74	**0.81**	0.97	0.70	**0.81**	0.89	0.74	**0.81**	0.97	0.79	**0.81**	0.89	0.74
	word	**0.81**	0.68	0.91	**0.80**	0.89	0.72	**0.81**	1.00	0.68	**0.85**	0.93	0.79	**0.80**	0.89	0.72
fastText	文字 3-gram	0.77	0.91	0.66	0.77	0.91	0.66	0.77	0.91	0.66	**0.83**	0.90	0.77	0.74	0.82	0.68
	word	0.75	0.94	0.63	0.75	0.88	0.65	0.76	0.97	0.63	**0.89**	0.97	0.83	0.75	0.88	0.65

ピペアを重複レシピペアと同数になるよう無作為に抽出する．評価には，leave-one-out 交差検証を用い，再現率，精度，F 値の観点から評価を行う．また，グリッドサーチを行い，ハイパーパラメータの最適化を行った．

表 2 に重複レシピペアの分類結果を示す．F 値に着目すると，すべてにおいて 0.7 を上回っており，素性に基づくレシピの分類が行えている．再現率と精度に着目すると，再現率が精度を上回っているケースが多い．これは，非重複レシピに誤分類される重複レシピが少ない一方，重複レシピと誤分類される非重複レシピが多いことを示している．重複レシピはレシピサイトのサービス品質の低下に繋がる恐れがあるため，再現率を重視した重複レシピの検出が求められる．よって，多くの手法で重複レシピを正しく分類できている点で評価できる．

また，多くの結果が SGNS を分散表現学習アルゴリズムとしたときのほうが良くなっている．これは fastText が分散表現の学習の際に subword 分割を行っていることによって，SGNS を分散表現学習手法としたときには距離があった単語，文字 3-gram についても，subword の一致によって，距離が近く算出され，重複レシピペアと非重複レシピペアの素性間で差が小さくなったために，分類が困難になったのだと考えられる．

本研究の貢献として，どの単語，文字 3-gram がどう置き換わっているかといった重複レシピと判断するための根拠を明示できるようになった点が挙げられる．例えば,「豚肉に火が通ったら火を止めます」「具材に火が通ったら火を止めます」という 2 文があった際,「豚肉」が「具材」に置き換わっていることがわかる．これを数値化して示せるようになったことで，重複の判定根拠としてユーザに提示しやすくなったと考えられる．

6 おわりに

本研究では，WMD を文字 3-gram の分散表現へと応用した手法を提案した．実験の結果，提案手法を用いると比較手法では検出できなかった一方のレシピに誤字や脱字が含まれている重複レシピについても検出できることが明らかになった．また，WMD を用いることで単語，文字 3-gram の置き換えがどのように行われているのかを明示的に示せるようになった．

一方，比較手法で検出できた重複レシピの中でも提案手法では検出できなかったものがあり，これらには語順の入れ替えや文書単位での類似した表現への言い換えが見られた．今後の課題として，実験で明らかになった提案手法の弱点を補えるようアルゴリズムの改善に取り組むことが挙げられる．

謝辞

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た．

参考文献

[1] Kusner, M. J., Sun, Y., Kolkin, N. I., et al.: From Word Embeddings To Document Distances, Proc. of the 32nd International Conference on International Conference on Machine Learning (ICML 2015), pp. 957-966, 2015.

[2] 小高知宏, 村田哲也, 高建斌ほか: 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, Vol. 86, No. 9, pp. 702-705, 2003.

[3] Tomas, M., Ilya, S., Kai, C., et al.: Distributed Representations of Words and Phrases and Their Compositionality, Proc. of the 26th International Conference on Neural Information Processing Systems (NIPS 2013), pp. 3111-3119, 2013.

[4] Bojanowski, P., Grave, E., Joulin, A., et al.: Enriching Word Vectors with Subword Information, Transactions of the Association for Computational Linguistics, Vol. 5, pp. 135-146, 2017.

[5] Oguni, M., Seki, Y., Shimada, R., et al.: Method for detecting near-duplicate recipe creators based on cooking instructions and food images, Proc. of the 9th Workshop on Multimedia for Cooking and Eating Activities (CEA 2017) in conjunction with the 2017 InternationalJoint Conference on Artificial Intelligence (IJCAI 2017), pp. 49-54, 2017.

[6] 小邦将輝, Nio, L., 平手勇宇ほか: 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研究報告, vol. 118, no. 278, pp. 19-24, 2018.

[7] Mekala, D., Gupta, V., Paranjape, B. et al.: SCDV : Sparse Composite Document Vectors Using Soft Clustering Over Distributional Representations, Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017), pp. 659-669, 2017.

[8] Cohen, J.: A Coefficient of Agreement for Nominal Scales, Educational and Psychological Measurement, Vol. 20, No. 1, 1960.

[9] Landis, J. R. and Koch, G. G.: The Measurement of Observer Agreement for Categorical Data, Biometrics, Vol. 33, No. 1, pp. 159-174, 1977.

[10] Yen, S. J. and Lee, Y. S.: Under-Sampling Approaches for Improving Prediction of the Minority Class in an Imbalanced Dataset, Proc. of the 2006 International conference on intelligent computing (ICIC 2006), pp. 731-740, 2006.