RecipeSTS: レシピのための類似性評価

山口泰弘　深澤祐援　原島純 *
クックパッド株式会社
{yasuhiro-yamaguchi,yusuke-fukasawa,jun-haranshima}@cookpad.com

概要

STS (Semantic Textual Similarity) は異なる 2 つのテキストの意味的な類似性を評価するための基本的なタスクである．レシピは自然言語で書かれる文書の一種であるものの，既存の STS データセットが対象とするテキストとは異なる性質を多く有するため，既存の STS ベンチマークで高い性能を達成するモデルがレシピデータにおいても高い性能を示すかどうかは明らかでない．本研究では日本語のレシピタイトルの意味的な類似性を評価するために新たなデータセットを作成し，学習済み言語モデルのレシピにおける性能評価を行った．

1　はじめに

BERT [1] や T5 [2] をはじめとした言語モデルの性能評価において，テキストの意味的な類似性を反映した振る舞いが見られるかどうかは重要な観点のひとつである．STS (Semantic Textual Similarity) [3] は自然言語で書かれた 2 つの文がどの程度意味的に似ているかを定量的に評価するタスクであり，現在では言語モデルの性能を測るためのベンチマークタスクとして広く利用されている．

レシピは自然言語で書かれる文書の一種であり，その意味的な類似性を適切に捉えることはレシピの検索や推薦といった応用において重要な役割を果たす．しかし，レシピは後述するように一般的な文書と異なる特有の性質をもつため，既存の STS ベンチマークで高い性能を達成するモデルがレシピの類似性を適切に扱うことができるかどうかについては議論の余地がある．

レシピは通常，タイトル・材料欄・調理手順・写真といった複数の要素からなる構造をもち，各要素は互いの情報を参照している．表 1 に Cookpad Recipe Dataset [4] に収録されているレシピの例を示す．この例では材料欄に記載されているいくつかの食材に記号 (“☆”) が割り当てられ，それらは手順 (3) のテキスト中で参照されていることがわかる．

タイトル	
和風な♪もやしの和え物	
材料欄	
もやし	1 袋
きゅうり	1/2 本
人参	1/2 本
カニカマ	3 本
☆オリーブオイル	ひとまわし
☆めんつゆ	ひとまわし
☆塩コショウ	少々
☆うまみ調味料	少々
手順	
1. もやしは茹でて水で冷ます	
2. きゅうり、人参は千切りにして塩をまぶし水気を切る。カニカマはさいておく。	
3. 器に 1 と 2 と☆の調味料を入れ混ぜたら完成〜♪	

図 1: レシピの例

レシピに記載されるテキストについても特有のスタイルが見られる．例えば，タイトルは名詞 (料理名) で終わる場合が多い，材料欄は食材名のほかに分量も併記される，手順では主語の省略・時間や順序に関する記載が多い，といった特徴が挙げられる．

本研究ではレシピの類似性評価に向けた取り組みの第一歩として，日本語で書かれたレシピのタイトルを対象にその類似性を評価するためのデータセットを作成し，学習済み言語モデルによる評価を行った．本稿では，データセットの作成プロセス，使用したアノテーションの方法，および複数の学習済みモデルを用いた評価結果について詳述する．また，実験の結果からレシピテキストの類似性を評価する際の課題と可能性について考察し，今後の研究の方向性を提示する．

2　関連研究

テキスト同士の意味的な類似性を評価するタスクとしては STS [3] がよく知られており，多くの基盤モデルの性能評価に利用されている．STS データセットは異なる 2 つの文のペアと，その類似度を示

* 現在は LINE ヤフー株式会社

す 0 ~ 5 のスコアからなる．日本語における文の類似性評価の取り組みとしては JGLUE [5] のサブタスクである JSTS がある．これらのタスクは一般的な文単体を評価対象としており，レシピのように特殊なドメインのテキストや構造化された文書の類似性を評価することを目的としたものではない．文書レベルの類似性を評価する手法として，Chen ら [6] は多言語のニュース記事を対象にしたタスクを提案した．この研究では，時間・場所・文体など複数の観点で文書同士の類似性を評価している．

大規模なレシピデータセットの事例として，日本語では Cookpad Recipe Dataset [4]，英語では Recipe1M+ [7] などがある．これらは材料欄や手順といった複数の要素からなる構造化されたレシピを収録している．また，Marin ら [7] はレシピの材料欄・手順と画像のベクトル表現を学習する手法を提案し，画像-テキスト間のマルチモーダルな検索タスクでその性能を評価した．

レシピのための基盤モデル開発に関する取り組みとしては，RECIPTOR [8] や RecipeGPT [9] が挙げられる．RECIPTOR は知識ベースを活用してレシピのベクトル表現を獲得する手法であり，RecipeGPT はレシピの一部から残りの情報を生成するモデルである．いずれの研究においてもモデルの性能は分類や生成といったタスクで評価されていて，異なるレシピ同士の意味的な類似性に関する定量的な評価は行われていない．

3 データセットの作成

RecipeSTS は STS と同様に，2 つの異なるレシピに対して意味的な類似性を示すスコアを割り当てるタスクである．この章ではレシピタイトルの類似性を評価するために作成したデータセットとその詳細について述べる．

3.1 レシピデータの収集

RecipeSTS データセットの作成にあたり，172 万品のレシピデータを収録した Cookpad Recipe Dataset [4] からレシピを抽出した．タイトルの重複を除いて 1,000 件のレシピからなる 500 個のペアを作成し，これらに対して人手によるアノテーションを行った．

無作為にペアを作るとほとんどの場合で低い類似性となるため，Cer ら [3] の手法に倣い，370 万件のレシピタイトルで事前に学習した fastText [10] を用いてタイトルのベクトルを作成し，そのコサイン類似度が高いものを優先的にペアとして選択した．より具体的には，タイトルの fastText ベクトルで近似近傍探索を行い得られた上位 8 件の中から無作為に選択したものをペアとした．

表 1: 類似性の評価基準と事例

類似度	基準 / 事例
5	２つのレシピが完全に一致している
	ジャガイモの冷たいスープ 冷たいじゃがいものスープ
4	ほとんど同じレシピだがそれほど重要でない違いがある
	ささみの梅肉あえ 鶏肉の梅肉あえ
3	似たレシピだが重要な情報が異なる・失われている
	たたききゅうりのしょうがあえ たたききゅうりの梅あえ
2	異なるレシピだが複数の共通の性質を持っている
	焼き肉のタレでおいしい鶏の唐揚げ 鶏肉の焼肉のタレで照る焼き
1	異なるレシピだが共通のトピックや語彙を持っている
	甘くないバナナブレッド 甘くないモカコーヒー
0	2 つのレシピは完全に異なっている
	かも鍋 圧力鍋でとりももホクホク♪親子丼

3.2 アノテーション

作成したレシピタイトルのペアに対して，表 1 の基準で人手による類似度のアノテーションを行った．クラウドソーシングサービスを利用し，ひとつのペアに対して 5 人の作業者が 0 ~ 5 の整数値を類似度として割り当てるよう依頼した．

各ペアに割り当てる類似度は，5 人の作業者が割り当てた類似度の平均値とした．作成したデータセットの詳細な統計情報を付録 A に記す。

4 実験

4.1 評価手法

作成したデータセットを元に，事前学習済みの BERT・T5 から得られた埋め込み表現がどの程度レシピの類似性を反映しているかを評価した．BERT と T5 のエンコーダの最終層を入力系列に渡って平均したベクトルをレシピタイトルの埋め込み表現として利用し，タイトル同士のコサイン類似度とアノテーションされた類似度の Spearman 相関係数 (ρ) を

表 2: RecipeSTS の評価に利用した言語モデル

モデル	パラメータ数	語彙数
BERT[a]	110M	32.7K
BERT-char[b]	110M	7.03K
T5[c]	60M	32.1K

[a] https://huggingface.co/cl-tohoku/bert-base-japanese-v3
[b] https://huggingface.co/cl-tohoku/bert-base-japanese-char-v3
[c] https://huggingface.co/retrieva-jp/t5-small-medium

表 3: RecipeSTS/JSTS による評価

モデル	JSTS (ρ)	RecipeSTS (ρ)
fastText	0.500	0.564
BERT	**0.765**	0.554
+ fine-tuning	0.702	**0.692**
BERT-char	0.725	0.585
+ fine-tuning	0.716	0.641
T5 (encoder)	0.437	0.321
+ fine-tuning	0.680	0.670

評価指標に採用した．

4.2 比較対象のモデル

表 2 に評価に利用した学習済み言語モデルの詳細を示す．これらのモデルはそれぞれ大規模な日本語のデータセットを用いて事前に学習されたものである．ここでは作成した全てのデータを評価に利用するため，各モデルに対して RecipeSTS に最適化するための追加の学習は行わず，事前学習によって得られた表現のみを利用した．

レシピドメインのテキストで学習された言語モデルについて評価を行うために，Cookpad Recipe Dataset から抽出した 170 万レシピのテキストを用いて表 2 に示したそれぞれのモデルに対して追加学習を行い，新たな言語モデルを作成した．

レシピのタイトル，材料欄，手順をあらかじめ定めた形式でまとめてひとつの文書とし，言語モデルの学習に利用した．BERT・BERT-char モデルについては Whole Word Masking によりランダムにマスクしたトークンの復元を学習し，T5 では材料・分量・手順の単位でランダムに欠落させたスパンを復元する Text-to-Text タスクを学習した．各モデルのより詳細な学習手法は付録 B に記す．

5 結果と考察

5.1 レシピの類似性評価

表 3 に RecipeSTS による各モデルの埋め込み表現を評価した結果を示す．最上部の fastText はベースラインとして 3.1 節でレシピタイトルのペアの作成に利用したレシピで事前学習したモデルによる評価の結果を示している．また，+ fine-tuning は各言語モデルに対してレシピデータを用いて追加学習を施したものを表す．

BERT をレシピデータで追加学習したモデルが最も高いスコアとなった．また，各モデルについて追加学習の前後のスコアを比較すると，いずれのモデルにおいても追加学習後のモデルで RecipeSTS スコアの改善が見られた．特に T5 においては他のモデルに比べてより大きな改善となった．

アノテーションされた類似度と埋め込み表現のコサイン類似度の分布を図 2 に示す．BERT における追加学習前後の分布の変化を見ると，事前学習前のモデルでは低い類似度がアノテーションされた事例においても比較的高いコサイン類似度を示す傾向があることがわかる．BERT と T5 についての追加学習後の分布の変化を比べると，BERT では低い類似度がアノテーションされた事例のコサイン類似度がより低くなり， T5 では高い類似度がアノテーションされた事例でコサイン類似度がより高くなる傾向が見られた．

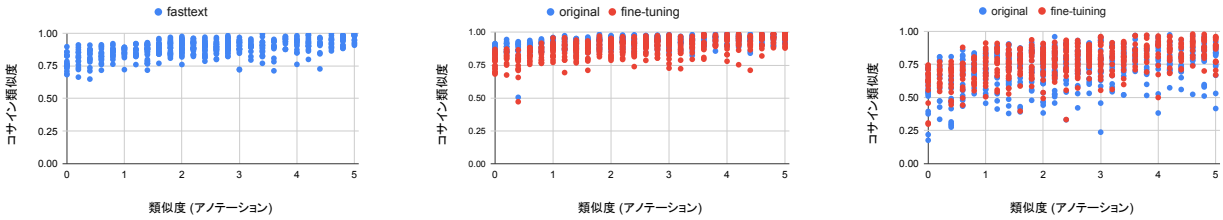
5.2 エラー分析

表 4 に RecipeSTS から選択したいくつかの事例と，それらに対して各モデルから得られたコサイン類似度を示す．

事例 (a) は表記の異なる同じ意味のテキストのペアであるものの，BERT・T5 のどちらにおいても追加学習前のモデルは比較的小さいコサイン類似度を示している．また，事例 (d) は全く異なる意味のペアであるが，いずれのモデルにおいても比較的大きいコサイン類似度となった．これらの事例は追加学習を通してコサイン類似度に改善が見られた．In-domain のデータを用いた学習により異表記や表層的な類似性に対してよりロバストな表現が得られるようになったと考えられる．

事例 (b) はほとんど同一の意味を持つタイトルのペアと考えられるが，各モデルのコサイン類似度は比較的小さい値を示している．さらに，追加学習後の結果と比較すると BERT・T5 の両方においてコサイン類似度が低下しており，改悪したことがわかる．“らぶりぃ” のようにそれほど一般的でない低頻度な表現を含む事例においては，類似性を正しく表現できない傾向が見られた．

事例 (c) はどちらも “オイスター炒め” であるものの，列挙されている食材が異なっている．コサイン



(a) fastText **(b)** BERT **(c)** T5

図 2: アノテーションとコサイン類似度の分布

表 4: 各モデルにおけるコサイン類似度

	事例	Gold	fastText	BERT		T5	
				original	fine-tuning	original	fine-tuning
(a)	あさりとにらの炒め物 にらとあさりのいためもの	4.4	0.926	0.855	0.930	0.697	0.775
(b)	らぶりぃトマトサラダ トマトサラダ	4.0	0.762	0.905	0.830	0.750	0.674
(c)	牛肉と白菜のオイスター炒め 豚肉とカブのオイスター炒め	2.4	0.957	0.978	0.966	0.897	0.890
(d)	にんじんのホットビスケット バナナのホットサンド	0.0	0.857	0.900	0.804	0.719	0.680

類似度はいずれのモデルにおいても比較的高い値を示した．この事例は調理方法と材料のどちらを重要と捉えるかによって類似性の評価が異なる場合があると考えられる．今回作成したアノテーションの基準 (表 1) ではどのような観点で類似性を評価するかを作業者の判断に委ねたが，今後の研究においてはレシピを比較する観点や状況を指定するなど，より具体的な評価を行うことも考えられる．

5.3 JSTS による評価

RecipeSTS の評価に利用したモデルが，より一般的なテキストにおいてどの程度意味的な類似性を考慮できているかを確かめるために JSTS データセットによる評価も合わせて行なった．JSTS の検証データによるスコアは表 3 に RecipeSTS の結果と併記している．

BERT の結果をみると，レシピを用いた追加学習によって RecipeSTS のスコアは上昇したが， JSTS のスコアは低下することがわかった．一方，T5 の結果をみるとレシピによる追加学習を行うことで RecipeSTS だけでなく JSTS においてもスコアの上昇が確認された．BERT の学習ではランダムにマスクされた単語を予測するのに対して，T5 の学習では材料や手順などレシピの意味や構造を考慮したスパンを予測している．こうした学習手法の違いが追加学習による JSTS スコアの変化に影響した可能性がある．この結果から，今回 T5 の追加学習に利用した手法はより一般的なドメインの言語モデルの学習においても有用であることが示唆された．

6 おわりに

本研究では日本語で書かれたレシピの意味的な類似性を評価するためのデータセットを作成し，学習済み言語モデルによる評価を行なった．実験の結果，一般的なドメインの文書で学習された言語モデルでは意味的な類似性を十分表現できない事例や，レシピの類似性評価における課題が示唆された．

よりレシピ特有の性質を考慮した類似性評価にむけて，今後は以下のような課題に取り組みたい:

レシピ構造の考慮 レシピを扱う基盤モデルの開発においては，その性能を正しく把握するためにレシピを構成する各要素とそれらの間の関係性を考慮した，よりレシピ全体の類似性を扱うデータセットによる評価が重要と考えられる．また，データセットだけでなく事前学習においてもレシピの構造的な特徴を考慮した手法を用いることで，より良いレシピ表現が得られる可能性がある．

多面的な類似性評価 5.2 節で述べたように，レシピの類似性評価は着目する観点によって変化すると考えられる．Chen ら [6] の取り組みと同様に，レシピの類似性についても調理方法・使用する食材・風味など複数の側面から評価するタスクを設計することで，レシピを扱うモデルの性質をより詳細に分析できると考えられる．

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2019)**, Vol. 1. Association for Computational Linguistics, June 2019.

[2] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **CoRR**, Vol. abs/1910.10683, , 2019.

[3] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens, editors, **Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)**. Association for Computational Linguistics, August 2017.

[4] Jun Harashima, Michiaki Ariga, Kenta Murata, and Masayuki Ioki. A large-scale recipe and meal data collection as infrastructure for food research. In **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)**, 2016.

[5] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**. European Language Resources Association, June 2022.

[6] Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, and Mattia Samory. SemEval-2022 task 8: Multilingual news article similarity. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan, editors, **Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval 2022)**. Association for Computational Linguistics, July 2022.

[7] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. **IEEE Trans. Pattern Anal. Mach. Intell.**, 2019.

[8] Diya Li and Mohammed J. Zaki. Reciptor: An effective pretrained model for recipe representation learning. In **Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2020)**. Association for Computing Machinery, 2020.

[9] Helena H. Lee, Ke Shu, Palakorn Achananuparp, Philips Kokoh Prasetyo, Yue Liu, Ee-Peng Lim, and Lav R. Varshney. Recipegpt: Generative pre-training based cooking recipe generation and evaluation system. In **Companion Proceedings of the Web Conference 2020 (WWW 2020)**. Association for Computing Machinery, 2020.

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

[11] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. **CoRR**, Vol. abs/1711.05101, , 2017.

A　RecipeSTS データセットの統計

作成した RecipeSTS データセットの統計情報を表 5，図 3 に示す。

表 5: RecipeSTS データセットの統計

ペア数	レシピ数	平均文字数	平均類似度
500	1000	11.4 ± 4.30	2.38 ± 1.37

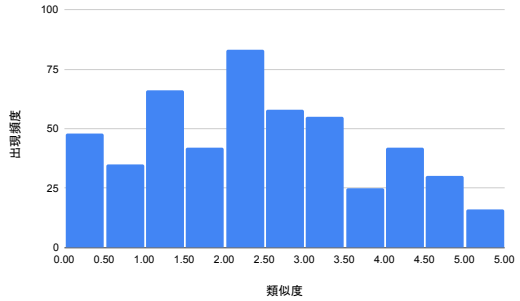


図 3: アノテーションされた類似度の頻度分布

B　レシピによる言語モデルの学習

レシピデータを使った言語モデルの学習に利用したデータセットと設定について詳述する．

B.1　データセット

Cookpad Recipe Dataset [4] から，タイトル，材料欄，手順がそれぞれからで空でない 170 万レシピを抽出し，そのうち 5,000 件を検証データとして利用した．

```
『{タイトル}』
概要: {レシピの概要}
分量: {レシピの分量}
材料:
  {材料名 1}: {材料 1 の分量}
  {材料名 2}: {材料 2 の分量}
  ...
手順:
  1. {手順 1}
  2. {手順 2}
  ...
```

図 4: レシピのフォーマット

レシピを単一のテキストとして表現するために，図 4 のフォーマットを用いて変換を行った．括弧 {} で囲まれたフィールドにそれぞれ対応するテキストが埋め込まれる．テキストのトークン数がモデルのコンテキストサイズを超える場合は分割してモデルに入力した．

BERT は Whole Word Masking を用いて 15% の確率でランダムにトークンをマスクし，マスクされたトークンを復元するよう学習した．一方，T5 では材料名・レシピの分量・材料の分量・手順をそれぞれ 10% の確率でランダムに欠落させ，欠損部分の予測を Text-to-Text タスクとして学習した．T5 の学習に利用した入出力テキストの例を表 6 に示す．なお，材料の分量においては出現頻度の高い “適量” のような曖昧な表記は予測の対象としないように設定した．

表 6: T5 の学習に利用したテキストの例

入力

```
『アボカドとグレープフルーツのサラダ』
概要: 材料 3 つ!簡単だけど美味しい♪
分量: <extra_id_0>
材料:
  アボカド: 1 個
  トマト: 1 個
  ピンクグレープフルーツ: <extra_id_1>
  エクストラバージンオイル: 大 2
  塩: 適量
手順:
  1. ピンクグレープフルーツは皮をむき、実をなるべく壊さないようにボールに取り出す。
  2. <extra_id_2>
  3. アボカドとトマトも 1 のボールに入れる。オリーブオイル、塩を入れ、軽くスプーンで 2~3 回混ぜて出来上がり!
```

出力

```
<extra_id_0>2〜3 人分<extra_id_1>1 個<extra_id_2>アボカドは皮と種をとり、一口大に切る。トマトも一口大に切る
```

B.2　学習方法と結果

バッチサイズ 32，最大ステップ数を 200,000 として AdamW [11] による学習を行った．学習率は 1×10^{-3} を設定し，線形スケジューラを用いて最初の 1,000 ステップを warmup としたあと徐々に減衰させた．また，1,000 ステップごとに検証データによる評価を行い，5,000 ステップ経過して改善が見られない場合に Early Stop するよう設定した．

学習の結果，BERT は 19,000 ステップほどで Early Stop した．一方，T5 は Early Stop することなく最大ステップ数に到達したためそこで学習を終了した．