DEIM Forum 2023 5a-5-5

商品説明に着目したレビュー文からの教師なしキーフレーズ抽出

ホンヘジン† 木村 優介†† 神田 悠斗† 吉丸 直希†† 波多野賢治†

† 同志社大学文化情報学部 〒 610-0394 京都府京田辺市多々羅都谷 1-3
†† 同志社大学大学院文化情報学研究科 〒 610-0394 京都府京田辺市多々羅都谷 1-3
E-mail: †{cgjf0040,cgjd0092}@mail4.doshisha.ac.jp, ††{kimura,yoshimaru}@mil.doshisha.ac.jp,
†††khatano@mail.doshisha.ac.jp

あらまし ユーザのレビューは有用な情報資源であり，他のユーザの意思決定や企業のマーケティング分析などに活用されている．しかし，インターネット上には膨大な量のレビューが存在しているため，機械的にレビューを分析し要約する必要がある．要約の一手法であるキーフレーズ抽出では，各レビュー文からそのレビュー文を代表するフレーズを抽出し，それらフレーズを集約することで商品ごとにレビューを要約する．しかし，レビューには配送やショップの対応などの商品とは関係ない内容が混在しており，商品とは関係ない観点は要約の際にノイズとなる可能性がある．そこで，本研究では商品説明を用いてレビュー文から商品と高い関係性を持つキーフレーズを抽出する手法を提案する．

キーワード キーフレーズ抽出，商品レビュー，商品説明

1 はじめに

EC サイトにおけるユーザの商品レビューは，他ユーザの意思決定や企業のマーケティング分析などにとって有用な情報資源である [1]．情報資源の活用を効率よく行うためには，機械的にレビュー内容を要約する必要がある．文書を要約する手法として，各文書を要約するキーフレーズ抽出と呼ばれる技術が挙げられる．

キーフレーズ抽出とは，文書からその文書を表す単語やフレーズを抽出する技術である [2]．キーフレーズ抽出には，教師あり・教師なしの手法が存在するが，文書にキーフレーズのラベルが付与されていないことが多いため，教師なしキーフレーズ抽出に対する需要は高い [3]．

キーフレーズ抽出には大きく分けて統計量ベース・グラフベース・深層学習ベースの 3 種類の手法が存在する [4]．統計量ベースの手法とは，キーフレーズの候補となる語（以後，候補語と記載）の出現頻度や文書頻度などの統計量を用いてキーフレーズを抽出する手法である [5], [6]．この手法は非常に高速だが，文書の意味を考慮することができないため，分散表現ベースの手法と比べると文書を要約する語・フレーズを抽出する精度は低い．

また，グラフベースの手法とは，候補語などで共起グラフを構築し，そのグラフから得られる特徴量を用いてキーフレーズらしさのスコア付けをする教師なしキーフレーズ抽出手法である．TextRank をはじめとするグラフベースの教師なしキーフレーズ抽出手法は，候補語をノードで表現し，候補語間の共起関係をグラフ上のエッジで表現することで，どれだけグラフ上で重要なノードかを候補語のキーフレーズらしさの値（以後，キースコアと記載）とした手法である [7], [8]．

最後に分散表現ベース手法とは，Word2Vec や BERT などから得られた候補語と文書の分散表現を用いて，候補語と文書の意味的類似度をキースコアとした教師なしキーフレーズ抽出手法である [9], [10]．EmbedRank や SIFRank などの分散表現ベースの手法は，それらの手法が用いている分散表現が共起関係も考慮しているため，グラフベースの手法よりも抽出精度が高い傾向にある [11], [12]．

分散表現ベースの手法は抽出精度が高いものの，レビューからキーフレーズを抽出する場合，商品そのもの（以後，商品属性と記載）とは関係のない候補語もキーフレーズとして抽出される可能性がある．例えば，商品属性とは関係のない候補語として，商品の配送や販売店の対応，他社と比較した際の他社製品の属性などが考えられる．レビューの要約を行う目的の一つとして商品分析があるが，既存のキーフレーズ抽出手法では商品属性と関係あるフレーズか否かを考慮する機構がないため，抽出されたフレーズが商品分析の際のノイズになる可能性がある．例えば，「注文したコーヒー豆の色はすごく可愛い我が家の猫ちゃんと色が似ています．コーヒー豆は酸味が少ないです.」と書かれたコーヒー豆に対するレビューはキーフレーズを抽出する場合，「すごく可愛い我が家の猫ちゃん」が抽出される．しかし，商品分析に必要である内容はコーヒー豆の酸味であり，抽出されたキーフレーズはコーヒー豆と関係ない内容であるため，商品分析の際のノイズになる．

そこで本研究では，商品属性と関係のあるキーフレーズを抽出するために，商品属性を含むと考えられる商品説明文書を用いた分散表現ベースの教師なしキーフレーズ抽出手法を提案する．商品説明文書とは，その商品の販売側が商品の情報や特徴を説明する文書であり，内容はユーザの意見とは関係なく内容は変わらない．本研究では，上記アイデアを分散表現ベースの教師なしキーフレーズ抽出手法である AttentionRank に組み込む [13]．AttentionRank では，各候補語のキースコアの算出は候補語とレビューの意味的類似度を用いていたが，提案手法

では候補語と商品説明文書の意味的類似度もキースコアの算出式に導入し，商品と関係あるキーフレーズだけをレビューから抽出できるようにする．

2 関連研究

高精度な分散表現ベースの教師なしキーフレーズ抽出手法として，Cross-Attention を用いて文間の関連性を考慮した AttentionRank と呼ばれる手法が存在する [13]．これまでの分散表現ベースのキーフレーズ抽出では，文書全体と候補語のそれぞれの分散表現のコサイン類似度をキースコアとしていた．しかし，文書には多種多様な意見が含まれているため，一文に一つの意見が含まれていると仮定すると，候補語がどの文と関連があるかを考慮する必要がある．

Cross-Attention は候補語がどの文と関係しているかを計算する一つの方法である．Self-Attention が文書に含まれるトークン同士でトークンの注目度を計算していたのに対し，Cross-Attention は文書に出現する候補語と文書内の各文の関係を元に，候補語に対する文の注目度を計算する方法である．AttentionRank は，Cross-Attention と Self-Attention を併用することで，候補語と文の意味的関係，候補語と文内それぞれに出現するトークン同士の意味的関係を考慮した候補語・文書ベクトルを構築することで，EmbedRank や SIFRank よりも文脈を考慮した上でキースコアを算出することができる．

3 提案手法

本研究では，商品との関連性が高い商品レビュー要約を行うために，商品説明文書を用いた教師なしキーフレーズ抽出手法を提案する．本研究の提案手法は AttentionRank に商品説明文書を考慮する機構を付与することで，候補語とレビューの意味的類似度だけではなく，候補語と商品説明文書との意味的類似度も考慮する点が AttentionRank との違いである．

本研究の提案手法を図 1 に示す．提案手法は，AttentionRank と同じ機構である図 1 の左側の，候補語とレビュー文書の意味的類似度を算出するステップと，新たに追加した図 1 の右側の候補語と商品説明文書の意味的類似度を算出するステップ，そして，前述の二つのステップで得られた 2 種類の意味的類似度の線形和を候補語のキーフレーズのキースコアとして算出する，図 1 の上部のステップの三つから構成される．

図 1 の右側の候補語と商品説明文書の意味的類似度を算出するステップだけを利用しても商品と関連性が高いキーフレーズ抽出が可能である．しかし，商品説明文書と意味的に類似している候補語だけがキーフレーズとして採択される可能性がある．したがって，本研究の提案手法では候補語とレビュー文書の意味的類似度，候補語と商品説明文書の意味的類似度を利用し，2 種類の意味的類似度の線形和を求め，ユーザの意見を重視しながら，商品と関連あるキーフレーズ抽出を行う．

本研究では，商品説明文書と候補語の関連性は AttentionRank を参考にし，Self-Attention と Cross-Attention を用いて算出する [13]．まず， 商品説明文書内の商品属性と考えられる名詞句を考慮するため，事前学習済みニューラル言語モデルに商品説明文書に出現する全名詞句をスペース区切りで入力する．また，レビュー文書と候補語をそれぞれ入力する．このとき，i はレビュー文書内の文番号を示し，k は商品説明文書に出現する名詞句の番号，c はそのトークンが候補語に属することを示す．n はレビュー文書の文内におけるトークン番号，m は候補語内のトークン番号，q は商品説明文に出現する各名詞句内のトークン番号である．これらの入力によって，事前学習済みニューラル言語モデルからトークン w の分散表現 e を得る．

3.1 第一ステップ

第一ステップでは，キーフレーズになることが多い名詞句（連接する一つ以上の名詞，もしくは名詞以外の品詞と名詞の組合せ）をキーフレーズの候補語として依存構造解析ツールで抽出し，候補語とレビュー文の分散表現をそれぞれ事前学習済み言語モデルから取得する．次に，それら分散表現のコサイン類似度をレビュー文書における候補語のキースコアとして用いる．詳細は以下である．

まず，事前学習済みニューラル言語モデルから抽出された分散表現 e を Cross-Attention に入力して他の文・名詞句との注意度 v を計算し，v を文・名詞句別の Self-Attention に入力して各文と名詞句の中での注意度 a_n^i を計算する．

$$a^i = \frac{\sum_{x=1}^{n} a_x^i}{n} \tag{1}$$

式 (1) では，レビュー文書の名詞句別の Self-Attention の出力値である a を各文書別の平均値を求め，Self-Attention に入力し，各名詞句がレビューをどのぐらい代表しているのかについての注目度 p_{review}^i を計算する．

$$p^{review} = \frac{\sum_{x=1}^{i} p_{review}^x}{i} \tag{2}$$

式 (2) では， 名詞句ごとの Self-Attention の出力値である p を各レビュー文書ごとの平均値を求める．式 (3) で c は候補語を意味し，m は候補語の個数である．

$$p^c = \frac{\sum_{x=1}^{m} a_x^c}{m} \tag{3}$$

式 (4) では，式 (2) から求めたレビュー文書の p^{review} と式 (3) で求めた p^c を利用してコサイン類似度を求める．

$$r_{review} = \frac{p^c \cdot p_{review}}{\| p^c \| \cdot \| p_{review} \|} \tag{4}$$

上の式から得られる値である r_{review} はキーフレーズを決定するキースコア式に使用してキーフレーズを決定する．

3.2 第二ステップ

第二ステップでは，商品属性を表す語で商品説明文書の分散表現を作成するために，説明文から商品属性になる可能性が高いと考えられる名詞句を候補語と同様の方法で抽出する．次に，候補語の分散表現と説明文に出現する全名詞句の分散表現のコサイン類似度を算出し，その値を候補語と商品説明文書の関連度として用いる．第二ステップは第一ステップと同様の計算方

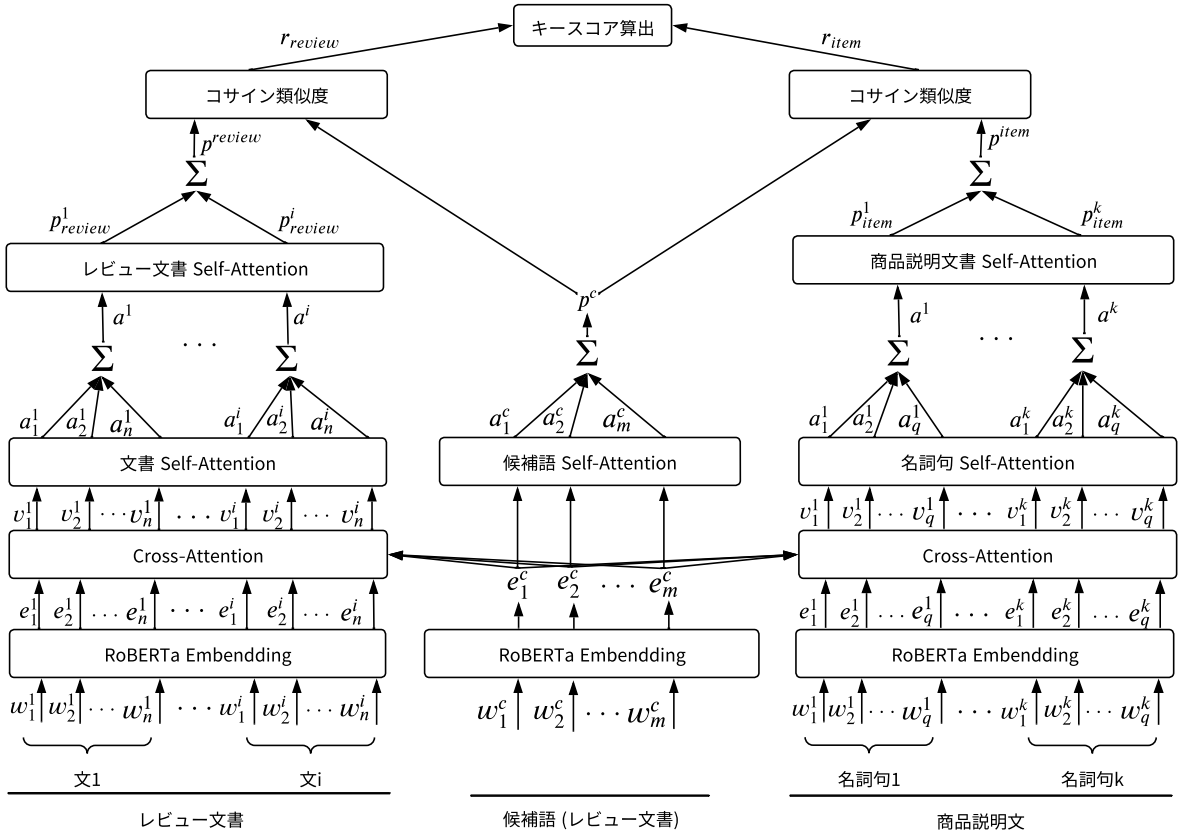

図 1　提案手法のアーキテクチャ

法であり，第二ステップの詳細は以下の式になる．

$$a^k = \frac{\sum_{x=1}^{q} a^k_x}{q} \tag{5}$$

$$p^{item} = \frac{\sum_{x=1}^{k} p^x_{item}}{k} \tag{6}$$

$$r_{item} = \frac{p^c \cdot p_{item}}{\| p^c \| \cdot \| p_{item} \|} \tag{7}$$

式 (7) から得られる値である r_{item} もキーフレーズを決定するキースコア式に使用してキーフレーズを決定する．

3.3　第三ステップ

第三ステップでは，第一ステップで得られたレビュー文書における候補語のキースコアと第二ステップで得られた候補語と商品説明文書の関連度を利用して，商品属性を考慮した候補語のキースコアを算出する．

キースコアにおいて α はハイパーパラメータであり，レビュー文と商品説明文書の間のコサイン類似度の平均を用いる．商品と関連のあるキーフレーズを抽出することも重要であるが，商品説明文書の内容と似たキーフレーズだけが抽出されることを防ぐためにレビュー文書と商品説明文書の両方を利用してキーフレーズ抽出を行う．つまり，レビュー文書と商品説明文書のどちらかに偏らないように式を構成した．

$$Key\text{-}Score = \alpha \times r_{item} + (1 - \alpha) \times r_{review} \tag{8}$$

4　評 価 実 験

評価実験では，本研究の提案手法の性能を検証するために，楽天グループ株式会社による楽天市場のデータセットから，レビューの長さが平均以上，かつ 42 件以上のレビューを持っている商品のレビューをランダムに 210 件抽出し，その商品の説明文書のデータを抽出して評価実験を行った [14]．楽天市場のデータセットにはキーフレーズの正解データが存在しないため，各レビューに対して三人の作業者で正解データとなるキーフレーズのアノテーション作業を実施した．作業の内容としては，まず始めに作業者に商品説明文書を読んでもらい，その後各レビューを代表していると思われる名詞，形容詞，形容動詞に対してキーフレーズのラベルを 5 語まで付与してもらう．

楽天市場データセットの商品説明文には表データが含まれているので，表データ処理を行い，名詞句を入力してキーフレーズ抽出を行う．

作成したデータセットを用いて，提案手法によって抽出されたキースコア上位 $k\,(= 1, 2, 3)$ 個のキーフレーズ候補の中に正

表 1　実 験 結 果

上位 k 語	手法名	適合率	再現率	F 値
1 語	AttentionRank	0.502	0.061	0.108
	提案手法	**0.729**	**0.087**	**0.155**
2 語	AttentionRank	0.467	0.114	0.183
	提案手法	**0.695**	**0.172**	**0.276**
3 語	AttentionRank	0.479	0.177	0.259
	提案手法	**0.634**	**0.240**	**0.348**

解データがどのぐらい含まれているかを確認した．さらに，既存手法と提案手法から得られたキーフレーズの上位 k 語までの適合率 Pre，再現率 Rec，F 値の平均値を求めた．式 (9) のうち，TP は上位 k 語の中に含まれる正しいキーフレーズの数であり，FP は上位 k 語の中に含まれているが，正解データではないキーフレーズの数である．また，FN は正解データであるが，上位 k 語の中に含まれていないキーフレーズの数である．

$$Pre = \frac{TP}{TP + FP}, \quad Rec = \frac{TP}{TP + FN} \tag{9}$$

$$F = \frac{2 \cdot Pre \cdot Rec}{Pre + Rec} \tag{10}$$

今回の評価実験では，依存構造解析ツールとして GiNZA[1] を使用し，事前学習済み言語モデルとして早稲田大学河原研究室の RoBERTa-base[2] を利用した．また，候補語である名詞句抽出のために自然言語処理のライブラリである spaCy[3] を使用した．

表 1 は，評価実験の結果である．提案手法は既存手法と比べ，上位 1, 2, 3 語のすべての評価指標で精度の改善を確認できた．特に，適合率で既存手法より提案手法の方がはるかに高いことが分かる．これらを通じて，商品説明文書を利用したキーフレーズ抽出の有効性を示した．

5 考　　察

評価実験を通じて，提案手法を利用してレビューだけではなく，商品説明文書に着目した方がよりレビューを閲覧するユーザに役に立つキーフレーズ抽出が可能であることが明らかになった．表 2 と 表 3 は，AttentionRank 単体では正解できなかったが，提案手法では正解できたキーフレーズ抽出の例である．

表 2 は自動車バッテリー関連部品に対するレビューであり，表 3 はマットレスに対するレビューである．AttentionRank を見れば，抽出されたキーフレーズは二つのレビューとも商品と関連のない内容であり,「笑」のように重要ではないものも抽出されていることが分かる．このような語をキーフレーズのランキングから逐一除外することも考えられるが，その度にルールを作成し検出していくことは現実的ではない．したがって，不要な単語を削除する過程を追加することなく，重要な情報を抽出できると言う点で提案手法に優位性がある．

表 2　抽出されたキーフレーズの例 1

上位 k	AttentionRank	提案手法
1	3 分	バッテリー
2	老眼	車
3	到着	コンパクト
4	笑	充電
5	着	バッテリー上がり

表 3　抽出されたキーフレーズの例 2

上位 k	AttentionRank	提案手法
1	眠り	寝心地
2	大人	マットレス
3	結果	今まで
4	お尻	購入
5	笑	高反発

次に，全体的に商品説明文書に沿ったキーフレーズを抽出できているかを確認するために，提案手法と既存手法のキースコア上位 3 位の中に，配送や店舗と関連している名詞句がどれほど含まれているかを確認した．配送・店舗と関係あるキーフレーズを「注文番号，発送，翌日，店舗」の四つに設定し調査したところ，提案手法には 15 個 (約 3.5%)，既存手法には 28 個 (約 6.2%) 含まれており，提案手法を用いることでキーフレーズから商品と関係ない内容が消滅していることがわかる．

しかし，評価実験で利用した正解データには候補語の対象ではない形容詞単体や形容動詞単体が含まれていた．この現象は，提案・既存手法は候補語を名詞句としてしまったためである．そのため，名詞句であるキーフレーズの中にユーザが重要だと思う単語が含まれてない場合を考案する必要があると考えられる．

6 おわりに

本研究では，商品説明文書を用いて商品と関係する語をレビューから抽出する教師なしキーフレーズ抽出手法を提案した．既存手法との評価実験では，すべての評価指標で既存手法を上回る精度を達成した．その理由は，本研究の提案手法が，ノイズになる配送・店舗側の対応など商品と関連性が少ない名詞句をキーフレーズとして扱うことを回避できたことに依ると考えられる．

今後の課題として，商品レビューだけではなくホテルやレストランなどの複数種類のレビューに対して提案手法の性能を確認する．また，本研究の提案手法は商品説明に当たる文書を自分で作成すれば，EC サイトだけではなく，複数のトピックを

1：GiNZA, https://megagonlabs.github.io/ginza/, アクセス日：2022/12/26

2：roberta-base-japanese, https://huggingface.co/nlp-waseda/roberta-base-japanese, アクセス日：2022/12/26

3：spaCy, https://spacy.io/, アクセス日：2022/12/26

含む論文や本の要約に使える可能性があるため，今後は EC サイト以外のドメインに対して実験を行うことを考えている．

謝　　辞

本研究は日本学術振興会科学研究費助成事業基盤研究 (B) JP19H04218 の助成を受けて遂行された．また，本研究では、国立情報学研究所の IDR データセット提供サービスにより楽天グループ株式会社から提供を受けた「楽天データセット」（https://rit.rakuten.com/data_release/）を利用した．ここに記して謝意を表す．

文　　献

[1] Charles H. Schwepker. Customer-Oriented Selling: A Review, Extension, and Directions for Future Research. *The Journal of Personal Selling and Sales Management*, Vol. 23, No. 2, pp. 151–171, 2003.

[2] Takashi Tomokiyo and Matthew Hurst. A Language Model Approach to Keyphrase Extraction. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18*, MWE '03, p. 33–40, USA, 2003. Association for Computational Linguistics.

[3] Xinnian Liang, Shuangzhi Wu, Mu Li, and Zhoujun Li. Unsupervised Keyphrase Extraction by Jointly Modeling Local and Global Context. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 155–164, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[4] Luo Xiao Ding, Haoran. AGRank: Augmented Graph-based Unsupervised Keyphrase Extraction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 230–239, Online only, November 2022. Association for Computational Linguistics.

[5] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. YAKE! Keyword Extraction from Single Documents Using Multiple Local Features. *Information Sciences*, Vol. 509, No. C, p. 257–289, jan 2020.

[6] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. *Automatic Keyword Extraction from Individual Documents*, chapter 1, pp. 1–20. John Wiley & Sons, Ltd, 2010.

[7] Rada Mihalcea and Paul Tarau. TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[8] Florian Boudin. Unsupervised Keyphrase Extraction with Multipartite Graphs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 667–672, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[9] 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 - Volume 2*, NIPS'13, p. 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.

[10] 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: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

[11] Kamil Bennani-Smires, Claudiu Musat, Andreea Hossmann, Michael Baeriswyl, and Martin Jaggi. Simple Unsupervised Keyphrase Extraction using Sentence Embeddings. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pp. 221–229, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[12] Yi Sun, Hangping Qiu, Yu Zheng, Zhongwei Wang, and Chaoran Zhang. SIFRank: A New Baseline for Unsupervised Keyphrase Extraction Based on Pre-Trained Language Model. *IEEE Access*, Vol. 8, pp. 10896–10906, 2020.

[13] Haoran Ding and Xiao Luo. AttentionRank: Unsupervised Keyphrase Extraction using Self and Cross Attentions. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 1919–1928. Association for Computational Linguistics, 2021.

[14] 楽天グループ株式会社. 楽天データセット. 国立情報学研究所情報学研究データリポジトリ, 2014. https://doi.org/10.32130/idr.2.0.