Wikipedia における文の品質推定のための大規模データセット

安道健一郎 [1]　関根聡 [2]　小町守 [1]
[1] 東京都立大学　[2] 理研 AIP
ando-kenichiro@ed.tmu.ac.jp

概要

Wikipedia は誰でも編集できるという特性上，客観的な百科事典の記述として適切ではない文が大量に含まれており，それらは日々編集者によりマークアップされている．本研究では，その作業支援を目的に Wikipedia 内の文の品質推定を行うためのデータセットを構築し、その自動検出タスクを実施した．作成したデータセットは英語版 Wikipedia の全編集履歴から抽出された約 341 万文に、大きく 5 種類に分類された 153 の品質ラベルが付与され、ノイズなどの処理をしたものである．このデータを使った分類ラベルの自動検出実験は、編集者をアノテーターとした文品質検出実験と見ることができ，73-85 程度の F 値が得られ、有用性が示された．

1　はじめに

Wikipedia は誰でも編集できることで有名な巨大なオンライン百科事典である．しかし，Wikipedia の質は長い間論争の的となっており，それは自然言語処理（NLP）にとって非常に重要な問題である [1, 2, 3, 4] ．Wikipedia のテキストは NLP のデータセットに広く利用され，主要なリソースとなっているため，Wikipedia の品質が NLP に与える影響は大きい [5, 6, 7, 8].

実際，いたずらなどによる質の悪い編集も存在する．そのような編集はしばしば他の編集者によって Wikipedia テンプレート [1] でマークアップされ，修正される．そのような，ユーザーの編集が繰り返されることで Wikipedia の質は信頼性面も含めて向上し続けているが，全ての質の悪い編集文を限られた編集者でチェックし修正することは非現実的である．そのため，この問題を機械でサポートする試みが過去にいくつかなされている．代表的なものは Wikipedia の Bot [2] で，自動で間違った Wikipedia マークアップ書式を修正したり，廃止された機能を新しいものに置換したり，特に，荒らし的な編集を元に戻すというものもある．しかし，これらは素朴で表層的なエラーを対象としており，対象外のエラーに対する多次元的で詳細な評価は人間がチェックしている．その他のサポートの試みとしては，記事全体の品質を推定する研究や，ある特定の品質ラベルを識別しようとする研究も行われている [9, 10]．前者の bot は文単位でのきめ細かい評価ができず，後者は限られた質の悪い点のみに着目している．

そこで，我々はきめ細かく様々な側面から文の品質を推定するための大規模データセットを構築した．文の品質ラベルとして，編集者が各文に対して付与した Wikipedia インラインテンプレート [3] を利用している．対象ラベルをトークページに関するものなどを除いて人手で厳選し，ノイズの多い文をフィルタリングした結果，品質推定ラベルは合計 153 個，総文数は約 341 万文になった．事前学習モデルを用いた自動検出の実験では，引用を必要とする文，構文や意味の修正を必要とする文，命題に関する問題を持つ文は検出が困難であることがわかった．このデータセットは Wikipedia の編集者をアノテーターとして大規模に文品質評価を行なったと見ることもでき，NLP の他タスクにおいても有用であると思われる．構築されたデータセットは一般公開されており利用可能である [4]．

2　品質推定データセット

2.1　ソーステキスト

Wikipedia は Wiki markup というマークアップ言語で書かれており，それが MediaWiki というパーサー

1） https://en.wikipedia.org/wiki/Wikipedia:Template_index/Cleanup

2） https://en.wikipedia.org/wiki/Wikipedia:Bots

3） https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Inline_Templates

4） https://github.com/ken-ando/WikiSQE_dataset

表 1　5 種類に分類した Wikipedia の品質評価ラベル．それぞれ頻度上位 5 ラベルの中から特徴的な 4 ラベルを選び，その総数と説明を示している．

Group	Count	Description
Citation		
Citation needed	2,373,911	コンテンツを検証するために参考文献が必要.
Dead link	84,101	外部リンクが壊れている.
Not in citation given	35,278	参考文献がコンテンツを支持していない.
Original research?	69,449	参考文献が第三者によって検証されていない.
Syntactic or semantic revision		
Clarification needed	138,739	文が理解しにくいので明確化が必要.
Vague	13,373	曖昧な表現を含んでいる.
Check quotation syntax	1,272	引用の構文がガイドラインに反している.
Weasel words	1,055	逃げ口上が含まれている.
Information addition		
Who?	91,924	特定できない形で人などが言及されているので具体化が必要.
When?	72,920	時間の表現が曖昧で明確化が必要.
Pronunciation?	31,517	発音情報が追加で必要.
Which?	23,387	曖昧な形でモノへの言及が行われているので具体化が必要.
Disputed claim		
Dubious	45,920	参考文献が付与されてはいるが，疑わしい主張.
Neutrality disputed	8,465	バイアスがある主張.
Relevant?	3,494	文が記事に関係あるか，そもそも百科事典に適しているか疑問.
Disputed	2,433	編集者によって主張の真実性が議論されている.
Other		
Disambiguation needed	107,953	曖昧さ回避ページではなく個別ページにリンクする必要がある.
Sic	50,658	誤りが含まれているが，ソースからコピーしたものである.
Needs update	19,550	最新の情報に更新する必要がある.
Emphasis added	2,444	引用に後から強調が追加されている.
Total	3,417,909	

により HTML に変換されている．本研究は英語版の全編集履歴をターゲットにしているため，この HTML 化処理は非常に計算量が多く，計算リソースと時間を必要とする．そのため，先行研究 [11] ですでに HTML 化されているものを活用する．このデータは 2019 年 3 月 1 日以前の英語版 Wikipedia の全記事の全編集履歴を含んでいる．

2.2　対象ラベル

収集対象の品質評価ラベルは Wikipedia inline cleanup template [5] に含まれるものをベースとして使用する．このインラインテンプレート集は編集者に向けて，Wikipedia における文の品質の指摘をするための Wiki markup をまとめたものである．我々が持つソーステキストは HTML であるため，このインラインテンプレートを Wikipedia サンドボックスを用いて HTML 表記に変換し，品質評価ラベルリストを獲得した．

しかし，この状態ではまだラベルリストのカバレッジに問題がある．第一にベースとして使用した Wikipedia inline cleanup template は 2022 年のものであり，ソーステキストで用いられた 2019 年の Wiki markup とは相違がある．そのため，各インラインテンプレートのページにリダイレクトしている，既に廃止された Wiki markup ページを再帰的に取得し，サンドボックスを通すことで過去に遡った HTML 表記の品質評価ラベルを獲得した．第二の問題は MediaWiki パーサーの時間的相違である．ソースの HTML 表記は 2019 年時点の MediaWiki によって生成されたものであり，現在の MediaWiki の出力する HTML ラベルとは相違がある．つまり，サンドボックスで変換した HTML ラベルがソーステキスト内に存在しない可能性がある．この問題に対処するために，ソーステキストに含まれる高頻度の品質評価ラベルを人手でチェックし，リストにないものは新たに収集リストに追加した．

結果，153 種類の品質評価ラベルが得られた．その一部を表 1 に示す．全取得ラベルとその説明はデータセットの Web ページにて公開されている．

5）https://en.wikipedia.org/wiki/Category:Inline_cleanup_templates

表 2　Wikipedia の品質評価ラベルと文例.

Label	Sentence
Citation needed	According to Japanese records, the term kendo is coined in Japan on August 1, 1919.[citation needed]
Dead link	Player profile at LFChistory.net[dead link]
Clarification needed	It was later given to the county, and has a possibility of becoming County Road 23.[clarification needed]
Vague	Pisces is perhaps[vague] the first hit rock or pop album to feature the Moog.
Who?	However, many analysts[who?] are finding that as Google grows, the company is becoming more ”corporate”.
When?	Over the last thirty years,[when?] a debate has been ongoing whether a tiny number of Ukrainians settled in Canada before 1891.
Dubious	Some academic linguists believe the modern English Language is half-Romance influenced (the evident Norman French influences), thus can be classified a Romance language.[dubious]
Neutrality disputed	Streetball is a very popular game worldwide, and a fun way for young people to keep out of trouble and avoid problems such as juvenile crime and drugs.[neutrality disputed]
Disambiguation needed	Attala County, Mississippi: Attala is named for Attala [disambiguation needed], a fictional Native American heroine.
Sic	He also notably sung ’Digital Surviver[sic]’, theme of Akiyama Ryo from Digimon Tamers.

2.3　品質カテゴリ

分析のために 153 の品質評価ラベルをさらに抽象化し 5 種類に分類した.（表 1，3）

Citation には引用に関するラベルが 59 個属する．このラベルは Citation needed という文に引用が必要であるということを示すラベルが大多数であり，データセット全体に渡ってみても 68%を占める．

Syntactic or semantic revision は文法や意味的な改善が必要ということを示すグループであり，26 個のラベルが属する．Clarification needed という曖昧でわかりにくい文意を明確にするべきというラベルが多数存在する．

Information addition は文に何かしらの追加の情報が必要であることを示すグループである．最も多いラベルは Who?であり，特定の人物名が明記されていないことを示すラベルである．

Disputed claim は文の形態に問題はないが，その命題に問題があることを示すグループである．最も多いラベルは Dubious であり，編集者から見て，不自然な，真偽が怪しい情報が含まれていることを示す．

最後に **Other** はどのグループにも属さないラベルの集合である．最も多いラベルは Disambiguation needed であり，文内の Wikilink が曖昧さ回避ページにリンクされていて改善が必要な際にマークアップされる．次いで多いラベルは Sic である．これはソースに忠実なものの，一般的にみれば文法的に誤りである事例が含まれている．

このカテゴリはさらに，Wikipedia に依存するものとそうでないもので分けられる．Syntactic or semantic revision，Information addition，Disputed claim (**SID**) の 3 カテゴリは一般の文品質にも共通する特性を備えているため，後の実験で分離して扱う．SID には約半数の 78 種類のラベルが含まれている．

表 3　5 種類の文品質カテゴリの統計量．# L はラベル数 # T は単語数，PPL はパープレキシティを示している．

Group	Count	# L	Ave. # T	PPL
Citation	2,694,604	59	26.83	90.48
Syn or sem revision	160,350	26	27.01	110.53
Information addition	310,853	32	27.82	79.37
Disputed claim	70,202	20	28.30	82.22
Other	181,900	16	34.01	110.92

2.4　文抽出とフィルタリング

文抽出はソーステキストを pySBD [12] を用いて文分割し，2.2 節で取得された品質評価ラベルリストを持つ文を抽出することで行った．しかし，ソーステキストはこのままではセクションタイトルや非文などが含まれており，ノイジーである．そのため，極端に短い文，Wiki markup がそのまま残っている文，頭文字が小文字である文などをフィルタリングした．各文は引用マーカーが取り除かれて重複が削除され，最終的に 3,417,909 文が取得された．

2.5　データセットの分析

取得できた文例を表 2 に示す．ラベルによって，文，単語，節に付加されるものがある．例えば，Citation needed や Clarification needed，Dubious，Neutrality disputed などはその性質上，文や節など対象のスパンに付加されやすい．一方で，Who?,

When?，Sic などは単語に付加されることが多い．

表 3 に GPT-2 を用いて算出した各カテゴリのパープレキシティを示した．パープレキシティはカテゴリの特性をよく反映しており，Syntactic or semantic revision は文法，意味的に一般とは外れた用法が多いため，パープレキシティが高くなっている．Other は Sic が多く含まれているため，特にパープレキシティが高い．また，Disputed claim は命題的に特異的ではあるものの，パープレキシティは低く，表層にその特徴が現れない，判定が難しいグループであることが予想できる．Information addition は追加の情報が必要だが，文自体に問題はない性質があるため，パープレキシティは低かった．

3 実験

Wikipedia の問題のある文の自動検出を行うために，作成したデータセットを用いて実験を行う．

3.1 セットアップ

実験は開発データ，テストデータに十分なサイズを用意するため，カテゴリごとに実行した（十分サイズが大きい頻度上位 20 個のラベルの実験結果については Appendix B を参照）．実験のためのデータセットとして，前節までに作成したデータの品質ラベルを除去したものを正例とし，新しくラベルの付与されていない文をランダムで同じサイズにサンプルしたデータを負例とする．開発データとテストデータのために，正例と負例をそれぞれ 500 文ずつランダムに抽出し，結合して 1000 文としたものを 2 つ用意する．残りのデータを訓練データとして使用した．また，Citation needed については突出してラベル数が多いため，カテゴリの評価をなるべく公平に行うために 200,000 文までにダウンサンプルして使用した．加えて，SID とすべてのラベルを含む All を用意する．それらは単なる加重平均ではなく，属するカテゴリのラベルを全て連結し，シャッフルした単独のデータセットである．検出に用いるモデルは DeBERTa [13] と BERT [14]，RoBERTa [15] をファインチューニングしたものであり，全ての設定について 2 値分類である．

3.2 実験結果

自動検出の実験結果を表 4 に示す．全体的には平均的に 7-8 割程度の検出精度であったが，引用の問題，構文や意味の修正を必要とする文，命題に関する問題を持つ文は検出が比較的困難であることが分かった．引用の問題については引用元の文献を勘案しなければならないため，Wikipedia 本文だけでは特定が難しいものと思われる．構文や意味の修正を必要とする文と命題に関する問題を持つ文については，高次に意味的な要素を捉えなければならないため難しい．また，何らかの情報の付加が必要な文は付加する対象の表現に特徴があるため，高精度に検出できたものと思われる（When?は時間表現を対象とするなど）．モデル間の比較では，総合的に BERT が最も良いスコアを示したものの，モデル間の性能に顕著な違いは見られなかった．最後に，すべてのデータを用いた All がどのモデルにおいても最低の精度であった．一方 SID は精度が高く，その差を勘案すると，カテゴリのまとまりを無視した横断的なデータを用いて学習した弊害であると思われる．SID は AII と比べて似た文をうまくクラスタリングできているので，うまく学習できた．

表 4 自動検出の実験結果（F1 スコア）．

Group	DeBERTa	BERT	RoBERTa
Citation	73.0	**74.5**	73.9
Syn or sem revision	73.0	**73.8**	71.9
Information addition	85.2	**85.3**	83.3
Disputed claim	**74.3**	73.2	74.2
Other	**82.6**	80.6	81.3
SID	79.5	79.0	**80.0**
All	70.4	**72.3**	71.6

4 関連研究

過去の Wikipedia の編集者支援のための品質推定研究は主に記事単位で行われている [16]．一方で，文単位に着目した品質推定の研究も存在する．それらは引用が必要ということを示す Citation needed [9]，誇張した表現を示す Puffery や Peacock [10]，曖昧な言い回しを示す Weasel words [17] ラベルを対象に文を集めて分析している．本研究との詳細な比較を Appendix A に示す．先行研究は一般に公開されていないものも多く，Wikipedia の一部の品質ラベルを対象としており，全体を含んだものではない．

5 おわりに

本研究では Wikipedia のための文品質推定データセットを構築し，ラベルの目的ごとにカテゴリを作成した．自動検出実験では平均的に 7-8 割程度の精度で検出できることが分かった．

謝辞

本研究は 国立研究開発法人情報通信研究機構委託研究「自動翻訳の精度向上のためのマルチモーダル情報の外部制御可能なモデリングの研究開発」及び JSPS 科研費 JP20269633 の助成を受けて行われた.

研究の遂行にあたり，ご議論いただいた東京大学の渡邊晃一郎氏に深く感謝いたします.

参考文献

[1] Jim Giles. Internet Encyclopaedias Go Head to Head. **Nature**, Vol. 438, pp. 900–901, 2005.

[2] Encyclopaedia Britannica. Fatally Flawed: Refuting the Recent Study on Encyclopedic Accuracy by the Journal Nature. **Chicago, Estados Unidos: Encyclopaedia Britannica**, 2006.

[3] Editorial. Britannica Attacks. **Nature**, Vol. 440, p. 582, 2006.

[4] Thomas Chesney. An Empirical Examination of Wikipedia's Credibility. **First Monday**, Vol. 11, No. 11, Nov. 2006.

[5] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 2383–2392, November 2016.

[6] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a Large-scale Dataset for Fact Extraction and Verification. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 809–819, June 2018.

[7] Mahnaz Koupaee and William Yang Wang. WikiHow: A Large Scale Text Summarization Dataset, 2018.

[8] Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. ParaCrawl: Web-Scale Acquisition of Parallel Corpora. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4555–4567, 2020.

[9] Miriam Redi, Besnik Fetahu, Jonathan Morgan, and Dario Taraborelli. Citation Needed: A Taxonomy and Algorithmic Assessment of Wikipedia's Verifiability. In **Proceedings of the 28th International Conference on World Wide Web Companion**, pp. 1567–1578, 2019.

[10] Amanda Bertsch and Steven Bethard. Detection of Puffery on the English Wikipedia. In **Proceedings of the Seventh Workshop on Noisy User-generated Text**, pp. 329–333, November 2021.

[11] Blagoj Mitrevski, Tiziano Piccardi, and Robert West. WikiHist. html: English Wikipedia's Full Revision History in HTML Format. In **Proceedings of the International AAAI Conference on Web and Social Media**, Vol. 14, pp. 878–884, 2020.

[12] Nipun Sadvilkar and Mark Neumann. PySBD: Pragmatic Sentence Boundary Disambiguation. In **Proceedings of Second Workshop for NLP Open Source Software**, pp. 110–114, Online, November 2020.

[13] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In **Proceedings of The 8th International Conference on Learning Representations**, 2020.

[14] 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**, pp. 4171–4186, June 2019.

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019.

[16] KayYen Wong, Miriam Redi, and Diego Saez-Trumper. Wiki-Reliability: A Large Scale Dataset for Content Reliability on Wikipedia. In **Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval**, pp. 2437–2442, 2021.

[17] Viola Ganter and Michael Strube. Finding Hedges by Chasing Weasels: Hedge Detection Using Wikipedia Tags and Shallow Linguistic Features. In **Proceedings of the ACL-IJCNLP 2009 Conference Short Papers**, pp. 173–176, Suntec, Singapore, August 2009.

表 5 先行研究との比較表．“Public?” はデータの公開状況を表す．

Inline label	# Sents	Public?
Citation needed [9]	36,140	No
Weasel words [17]	500	No
Peacock, Puffery [10]	284	Yes

表 6 頻度上位 20 品質ラベルの結果．各ハイライトは Citaion，Syntactic or semantic revision，Information addition，Other，Disputed claim の上位カテゴリを表す．

Label	Count	F1
Citation needed	2,373,911	70.0
Clarification needed	138,739	74.8
Disambiguation needed	107,953	82.3
Who?	91,924	88.7
Dead link	84,101	77.9
When?	72,920	87.4
Original research?	69,449	89.9
Sic	50,658	92.9
Dubious	45,920	75.6
By whom?	41,588	92.9
Not in citation given	35,278	75.2
Pronunciation?	31,517	98.8
Attribution needed	27,322	94.3
Unreliable source?	25,369	74.1
Which?	23,387	86.0
Needs update	19,550	92.0
Verification needed	18,311	74.2
According to whom?	15,205	83.1
Vague	13,373	74.1
Neutrality disputed	8,465	82.7

A 先行研究との比較

Wikipedia の inline cleanup template を使用し，文の品質推定を行った先行研究との詳細な比較を表 5 に示す．

B 個別ラベルの自動検出結果

品質ラベルの頻度上位 20 ラベルについて DeBERTa で自動分類した結果を表 6 に示す．

C 実験の詳細

DeBERTa，BERT，RoBERTa はすべて Base_uncased モデルを用いた．エポックは最大 20 に設定し，F1 が最大であるベストエポックモデルをテストに使用した．最大入力系列長は 256，バッチサイズは 64、学習率は 1e-5 である．