論述リビジョンのためのメタ評価基盤

三田 雅人 [1]　坂口 慶祐 [2]　萩原 正人 [3,4]　水本 智也 [5,1]　鈴木 潤 [6,1]　乾 健太郎 [6,1]
[1] 理化学研究所　[2]Allen Institute for AI　[3]Earth Species Project
[4]Octanove Labs　[5] フューチャー株式会社　[6] 東北大学
masato.mita@riken.jp

概要

論述やエッセイの作文のように文書単位で行うリビジョンには，従来の文単位文法誤り訂正の研究範囲では捉えきれない論述全体の流れや一貫性といった要素が含まれる．また，文書単位のリビジョンは妥当な参照が多岐にわたることから高精度な参照なし評価尺度の実現が大きな課題となる．本研究では，自動論述リビジョンの実現に向けて，高精度な自動評価尺度の開発促進を目的としたメタ評価基盤を提案する．そして，大規模言語モデルを用いたベースライン自動評価尺度を用いた自動評価の現状と実現可能性を示す．

1　はじめに

論述やエッセイ等の作文において，リビジョンは重要な段階である．プロセスライティング教育学では，作文には，まず最初に文書単位で全体的な編集を行う *Revision* があり，その後，文または句単位での編集を行う *Editing*，最後に単語単位での細やかな編集を行う *Proofreading* の 3 段階があるとされている [1, 2].

これに対し，自然言語処理（NLP），特に文法誤り訂正（GEC）分野では単語単位を中心としたスペリングや文法誤りなどを対象とした局所的な編集（*Minimal edit*）[3, 4] から，句や文単位で流暢性を考慮した編集（*Fluency edit*）[5, 6] へと研究範囲を広げてきたといえる（図 1）．リビジョンには，従来の文単位 GEC の研究範囲では捉えきれない，論述全体の流れや一貫性・結束性といった評価項目に基づく様々な編集が含まれると考えられる．しかし，NLP におけるリビジョンに関する既存研究の多くは文単位を対象にしており [7, 8]，また論述やエッセイなどを文書単位で自動的にリビジョンを行うタスク（本稿では，**論述リビジョン**と呼ぶ）も存在しない．

論述リビジョンの設計を考えた場合，高精度な参照なし自動評価の実現が大きな課題となる．なぜなら文書単位のリビジョンの場合，文結合や文分割，文の並び替えなどといった文を跨いだ編集も含まれるため参照との正確な照合自体が困難であり，また妥当な参照も多岐にわたるため参照あり評価は現実的ではないからである（図 1 の実例を参照）．参照なし評価尺度を用いる場合，その評価尺度がどの程度人間の判断と相関があり信頼できるものかといった自動評価尺度の“評価（メタ評価）”が必要となる．

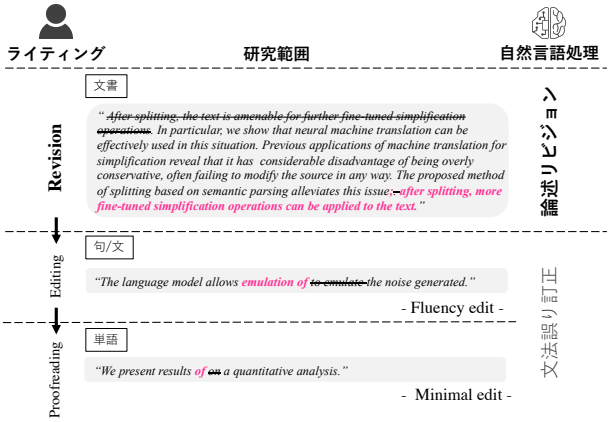

図 1　本研究の位置付け．図中の例は本研究で構築したデータセットに含まれていた実例である．

そこで本研究では論述リビジョンの実現に向けて，**評価項目毎に高精度な参照なし自動評価尺度の開発促進**を目的としたメタ評価基盤を提案する．具体的には，提案メタ評価基盤は論述リビジョンのためのデータセットである **T**ext **R**evision of **A**CL paper (**TETRA**)，およびメタ評価手法である **I**nstance-based **R**evision **C**lassification (**IRC**) の 2 つによって実現する [1)]．TETRA は，ACL 系論文に対して文書レベルのリビジョンをアノテーションしたデータセットであり，従来の局所的な編集タイプに加えて，文書単位での大域的な編集タイプにも対応可能なアノテー

1)　構築したデータセットは今後一般公開する予定である．

ションスキームに基づいて設計されている．IRCは，人間の専門家によるリビジョンに対して，編集事例毎にどの程度開発した参照なし評価尺度がリビジョンの良し悪しを判別できるかといった二値分類に基づくメタ評価手法であり，その精度が高い参照なし評価尺度が良い評価尺度ということになる．さらに，提案メタ評価基盤では評価項目毎の精度を評価および編集根拠を提示可能であるため，より良い評価尺度の開発に向けた透明性・解釈性の高い分析ができる．

本研究では，大規模言語モデルに基づくベースライン自動評価尺度（ラベル教師あり，なし）を用意し，それらが文書レベルのリビジョンをどの程度正確に判別することができるかについてのメタ評価を行い，文書レベルのリビジョンに対する自動評価の現状と実現可能性を示す．

2 TETRA

2.1 データセット設計

論述リビジョンのためのデータセットとして，どのように設計すべきかは自明ではない．そのため，本研究ではまず基本要件として次の 4 つを定めた:（1）段落単位で十分に長い文脈を含んでいる;（2）文法誤りは修正済みである;（3）ドメインは限定的である;（4）書き手の多様性を担保する．要件 1 は，論述リビジョンであるための必要条件ともいえる．要件 2 に関して，文法誤りは既存の GEC タスクで対象としていること，また本研究の狙いである文書単位での品質を向上させることを目的とした大域的かつ高次な編集は，事前により低次な文法誤りが修正されていなければ観察されにくいという仮定に基づいている．要件 3 に関して，リビジョンは暗黙的なドメインに依存した評価項目（rubric）の存在を前提としており，ドメインを制限しなければ妥当な編集候補が多くなり不良設定問題になるのを避けるためである．要件 4 に関して，リビジョンの書き手の属性は本質的に多様性があるためデータを収集するうえでも偏らないよう多様性を担保することが重要である．

上記の 4 つの要件を満たすために，次の方法論でデータの選定および収集を行う．まず，要件 1 から 3 を満たすために，本研究では ACL 系論文 の概要および導入節を元データとして採用する．概要および導入節は一般的に他の節よりも言説構造が重要であり，かつ数式などの非テキスト要素が入りにくいと考えられ，多様なリビジョンを収集しやすいと考えられる．アノテーション対象の論文を選択する際は，要件 4 を考慮して，次の 3 つの観点で多様性を担保する:（1）会議/ワークショップ;（2）学生/非学生;（3）母語話者/非母語話者．ここで，観点 2 および観点 3 について機械的に判別することは困難であるため，次の方法論で可能な限り恣意性を除外したうえで選択した．まず，ACL Anthology [2] の会議識別 ID を基に会議およびワークショップ論文をランダムサンプリングする[3]．次に，選択された論文を上記 3 観点の組み合わせから構成される全 8（2^3）クラスに人手で分類する [4]．各クラスにつき 1 本該当論文を割り当てるまでを 1 バッチとし（つまり，1 バッチにつき 8 論文を選定），本研究では結果としてこれを 8 バッチまで行い，合計 64 論文をアノテーション対象の論文として収集した．

2.2 アノテーション

本研究では英語母語話者，かつ英文校正の専門家である 3 名のアノテータによりアノテーションを実施した．具体的には，同じ文書を 3 名が独立的に編集し，各文書につき 3 つの参照を作成した．アノテーションとして，本研究では 1 段落を 1 文書とみなして段落単位でリビジョンを行うよう指示した．さらに，各編集に対してどういった観点で段落全体の品質が向上されるかについて編集タイプおよびその編集根拠を自由記述で書くよう指示を与えた．ここで，事前に定義したタイプ集合から選択させるのではなく自由記述とした理由は，文書単位リビジョン固有の言語現象として，どのような種類の編集が存在・観察されるのか自体自明でないからである．

次に，アノテーションしたデータを研究用途として利用しやすい形式にするために，XML を用いて人手でデータセットの構造化を行った．具体的には，元論文およびアノテータの識別 ID，節情報といったメタ情報に加え，編集事例毎に編集情報，編集タイプ及び編集根拠を付与した．TETRA の統計量を表 1 に示す．

2） https://aclanthology.org/
3） 本研究では，会議論文の識別 ID を P，ワークショップ論文の識別 ID を W とみなしてそれぞれ収集した．
4） 母語話者かどうかを判断する際，著者の母語を考慮するのは現実的に難しく，かつ差別的になる恐れがあるため，本研究では筆頭著者の所属の所在地を判断基準とした．

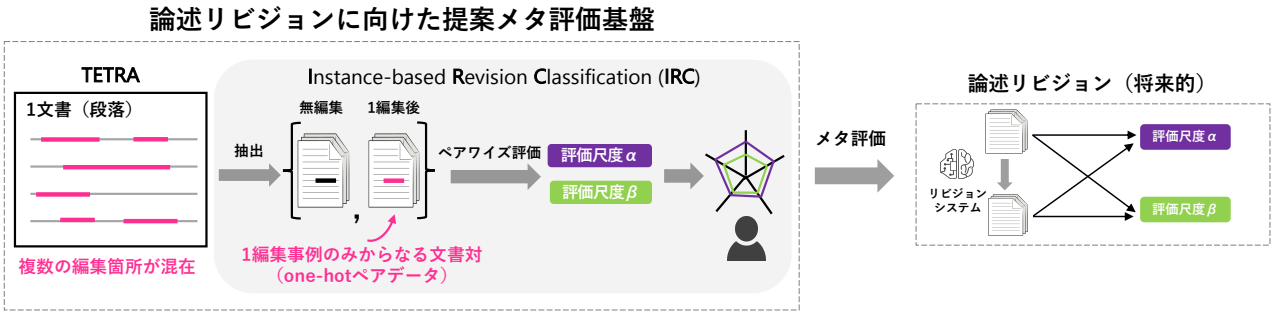

図 2　提案メタ評価基盤を用いたメタ評価の概要.

表 1　TETRA の統計量.

文書対数	386
平均編集文書割合（%）	87.9
リファレンス数	3

表 2　各評価項目の分布.

評価項目	編集タイプ（抜粋）	#	%
Grammaticality	grammar	81	22.1
Fluency	word choice, word order	42	11.4
Clarity	clarity	43	11.7
Style	style, tone	5	1.4
Readability	readability, punctuation	160	43.6
Redundancy	redundancy, conciseness	28	7.6
Consistency	consistency, flow	8	2.2

2.3　分析

構築した TETRA にどのような種類の編集が観察されるか，また従来の GEC では観察されない論述リビジョン固有の編集はどの程度含まれるかについて分析するために，一人のアノテータにおける 2 バッチ分のサンプルデータ（計 16 論文）に対して編集タイプの分布を算出した（表 2）．ここで，分析のしやすさを目的に各編集タイプを大分類として評価項目毎に人手でまとめた[5]．その結果，Grammaticality（文法性）に関する編集や Fluency（流暢性）に関する編集といったような従来の GEC に含まれるような編集も観察される一方で，Clarity（明瞭さ）・Style（スタイル）・Readability（読みやすさ）・Redundancy（冗長性）・Consistency（一貫性）といった論述リビジョン固有の編集も全体の 66.5%と大半は論述リビジョン固有の編集だった．

次に，アノテータ間でどのくらい編集が一致したかについて分析する．論述リビジョンにおいては，各編集のスパンの同定およびアノテータ間での対応を機械的に正確に取得すること自体が困難であるため，本研究では少量のサンプルデータに対して人手でアノテータ間の編集の対応を取り一致率を計算した．具体的には，ランダムに選んだ 3 本の論文に対して，3 人中 2 人以上が概ね同じ箇所を編集していたときの一致率を計算したところ，わずか 33.5%という結果となった．このことからも，論述リビジョンは妥当な参照が多岐にわたることがわかる．

3　提案メタ評価基盤

入力に多数の評価項目が含まれる論述リビジョンの評価方法として，絶対評価は難しいと考えられる．そのため，最も素直なメタ評価手法としては，人間の専門家によるリビジョン（gold revision）を用いた二値分類の Accuracy で評価を行うことが考えられる．具体的には，入力としてリビジョン前後の 2 つの文書が与えられたとき，どの程度 gold revision に対して評価尺度が正しく改善と判定できるかの二値分類であり，その精度が高い評価尺度が良い評価尺度ということになる．二値分類のようなペアワイズ比較は，絶対評価が難しい状況下のメタ評価手法として有効であることが先行研究でも示されている [9, 10]．しかし，表 2 に示されるような多数の評価項目に基づく多種多様な編集を一括りにして改善したかどうかの二値を提示するだけでは透明性や解釈性の高い分析が難しい．また，評価項目に応じて最適な評価尺度も異なると考えられるため包括的な評価では限界がある [11, 12]．

そこで本研究では，評価項目毎に高精度な評価尺度の開発促進を目的としたメタ評価基盤を提案する．図 2 に提案メタ評価基盤を用いたメタ評価の概要を示す．提案メタ評価基盤は，2 節で構築した TETRA とメタ評価手法，**I**nstance-based **R**evision **C**lassification (**IRC**) によって実現する．IRC では，複数の多種多様な編集が混在する文書に対して，1 文書につき 1 編集事例のみからなる文書対（本稿では，**one-hot ペアデータ**と呼ぶ）に変換して二値分類

5）　観察された編集の実例は付録 A に示す.

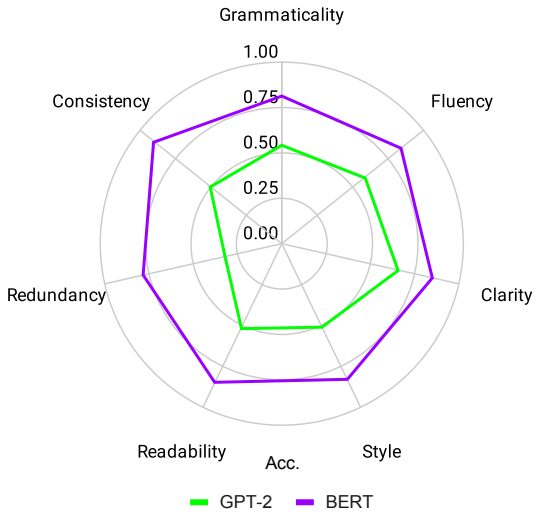

図 3　メタ評価結果（Accuracy）.

を行うことでメタ評価を行う．これにより，評価項目毎の精度および編集根拠を提示可能となるため，より良い評価尺度の開発に向けた透明性・解釈性の高い分析ができる．

4　実験

本節では，ベースラインとなる参照なし評価尺度を用意し，提案メタ評価基盤を用いたメタ評価のデモンストレーションを行う．そして，文書レベルのリビジョンに対する自動評価の現状と実現可能性を明らかにする．

4.1　実験設定

評価　提案メタ評価基盤でメタ評価を行うために one-hot ペアデータを構築する．具体的には，TETRA を論文単位で訓練: 評価用に 3（48 論文）:1（16 論文）に分割し，評価用データを one-hot ペアデータに変換した．ここで，TETRA に付与されている編集タイプは自由記述でアノテーションされている都合上，アノテータ間で異なるラベル集合となっている [6]．そこで本実験では，2.3 節の分析で用いたアノテータ（主アノテータ）のラベル集合を基準とし，残りの二人のアノテータの編集タイプを必要に応じて人手で主アノテータの編集タイプへと対応付けを行った．上記の手順に従い one-hot ペアデータを構築した結果，1368 文書対となった．

ベースライン評価尺度　本実験では，ベースラインとして 2 つの大規模言語モデル（GPT-2 [13], BERT [14]）に基づく参照なし評価尺度（二値分類器）を構築した [7]．GPT-2 を用いたラベル教師なし評価尺度は，入力の二文書それぞれの単語あたりの perplexity を比較し，その大小によって二値分類を行う．BERT を用いたラベル教師あり評価尺度は，TETRA の訓練用データ（全 868 文書対）に対し，半分はランダムにリビジョン前後を入れ替えて負例を作成し，それらを用いて二値分類問題として finetune を行った．なお，ベースラインの構築はいずれも `transformers` [15] の Pytorch 実装を用いた．

4.2　結果

図 3 に提案メタ評価基盤を用いたメタ評価結果を示す [8]．まず，図 3 に示されるように，本提案メタ評価基盤を用いると各評価尺度における評価項目毎の精度が評価可能である．これにより，ユーザは各評価尺度の得意・不得意を分析しながら評価項目毎に最適な評価尺度の開発に専念できる．また本実験結果から，ラベル教師なし評価尺度ではほとんど文書レベルのリビジョンを捉えることができないが，ラベル教師あり評価尺度においては，最も低い評価項目で 0.79 ポイント，最も高い評価項目では 0.90 ポイントの Accuracy を達成していることから，ある程度文書レベルのリビジョンを捉えることができていることがわかる．これはつまり，論述リビジョンという挑戦的な未開拓課題において自動評価の実現可能性が示されたといえる．

5　おわりに

本研究では，文法誤り訂正の次なる方向性として文書単位で自動的にリビジョンを行う論述リビジョンという新たな課題を提示し，論述リビジョンのための高精度な参照なし自動評価尺度の開発促進を目的としたメタ評価基盤を提案した．また，大規模言語モデルを用いたベースライン評価尺度を用意し，提案メタ評価基盤を用いたメタ評価のデモンストレーションを通して論述リビジョンにおける自動評価の実現可能性を示した．論述リビジョンは NLP 分野として重要かつ挑戦的な課題であり，本研究は論述リビジョンの実現に向けた最初の一歩となる．今後は，論述リビジョン自体を行うリビジョンモデルだけでなく，本提案メタ評価基盤を用いた自動評価尺度に関する研究に繋がることを期待する．

6）　例えば，語彙選択に関する編集に対してアノテータ間で word choice や word use など異なるラベルを用いていた．

7）　各評価尺度の詳細な実験設定は付録 B に示す．

8）　詳細な結果は付録 C に示す．

参考文献

[1] M. Buchman, R. Moore, L. Stern, and B. Feist. *Power Writing: Writing with Purpose*. No. No. 4. Pearson Education Canada, 2000.

[2] Anthony Seow. *The Writing Process and Process Writing*, p. 315–320. Cambridge University Press, 2002.

[3] Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. The CoNLL-2013 Shared Task on Grammatical Error Correction. In *Proceedings of the 17th Conference on Computational Natural Language Learning (CoNLL 2013): Shared Task*, pp. 1–12, 2013.

[4] Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL 2014): Shared Task*, pp. 1–14, 2014.

[5] Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Transactions of the Association for Computational Linguistics*, Vol. 4, pp. 169–182, 2016.

[6] Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pp. 229–234, 2017.

[7] Fan Zhang, Homa B. Hashemi, Rebecca Hwa, and Diane Litman. A corpus of annotated revisions for studying argumentative writing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1568–1578. Association for Computational Linguistics, 2017.

[8] Takumi Ito, Tatsuki Kuribayashi, Hayato Kobayashi, Ana Brassard, Masato Hagiwara, Jun Suzuki, and Kentaro Inui. Diamonds in the rough: Generating fluent sentences from early-stage drafts for academic writing assistance. In *Proceedings of the 12th International Conference on Natural Language Generation*, pp. 40–53. Association for Computational Linguistics, 2019.

[9] Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. Pairwise neural machine translation evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 805–814. Association for Computational Linguistics, 2015.

[10] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., 2017.

[11] Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A. Smith. Transparent human evaluation for image captioning. *arXiv* https://arxiv.org/abs/2111.08940, 2021.

[12] Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander R. Fabbri, Yejin Choi, and Noah A. Smith. Bidimensional leaderboards: Generate and evaluate language hand in hand. *arXiv* https://arxiv.org/abs/2112.04139, 2021.

[13] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[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, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

[15] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45. Association for Computational Linguistics, 2020.

[16] Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, 2015.

A　リビジョンの実例

表 3　評価項目に関連付けされたリビジョンの実例.

Readability　Redundancy　Style　Clarity

This paper presents empirical studies and closely corresponding theoretical models of a chart parser’ s performance while~~the performance of a chart parser~~ exhaustively parsing the Penn Treebank with the Treebank’s own context-free grammar (CFG)~~CFG grammar~~. We show how performance is dramatically affected by rule representation and tree transformations, but little by top-down vs. bottom-up strategies. We discuss grammatical saturation, provide an~~, including~~ analysis of the strongly connected components of the phrasal nonterminals in the Treebank, and model how, as sentence length increases, regions of the grammar are unlocked, increasing the effective grammar rule size ~~increases as regions of the grammar are unlocked,~~ and yielding super-cubic observed time behavior in some configurations.

Modeling relation paths provides ~~has offered~~ significant gains in embedding models for knowledge base (KB) completion. However, enumerating paths between two entities is very expensive, and existing approaches typically resort to approximation with a sampled subset. This problem is particularly acute when text is jointly modeled with KB relations and used to provide direct evidence for facts mentioned in it. In this paper, we propose the first exact dynamic programming algorithm , which enables efficient incorporation of all relation paths of bounded length, while modeling both relation types and intermediate nodes in the compositional path representations. We then conduct a theoretical analysis of the efficiency gain from the approach. Experiments on two datasets show that it addresses representational limitations in prior methods ~~approaches~~ and improves accuracy in KB completion.

B　ベースライン評価尺度の実験設定

GPT-2 を用いた評価尺度　GPT-2 を用いたラベル教師なし評価尺度は，入力の二文書それぞれの単語あたりの perplexity を比較し，その大小によって二値分類を行う．具体的には，リビジョン前の文書よりもリビジョン後の文書の perplexity の方が低い場合は改善（正例），高い場合は改悪（負例）とみなして二値分類を行う．

BERT を用いた評価尺度　BERT を用いたラベル教師あり評価尺度は，TETRA の訓練用データ（全 868 文書対）に対し，半分はランダムにリビジョン前後を入れ替えて負例を作成し，それらを用いて二値分類問題として finetune を行った．具体的には，入力として，“リビジョン前<SEP>リビジョン後” を改善（正例），“リビジョン後<SEP>リビジョン前” を改悪（負例）という形式で与え，BERT+線形分類レイヤによって finetune することで二値分類器を訓練した．なお，モデル訓練時のハイパーパラメータは表 4 に示す．

表 4　BERT を用いた評価モデル訓練時のハイパーパラメータ

ハイパーパラメータ	設定値
モデル	bert-base-uncased
最適化器	Adam [16]
学習率	2e-5
エポック数	10
バッチサイズ	32

C　メタ評価結果

表 5　メタ評価結果の詳細.

Acc.	Grammaticality	Fluency	Clarity	Simplicity	Readability	Monotonicity	Consistency
GPT-2	0.54	0.58	0.65	0.51	0.52	0.32	0.50
BERT	0.82	0.84	0.85	0.83	0.85	0.79	0.90