一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
DE2015-17(2015-08)

クラウドソーシングによる Wikipedia 編集者の質推定

鈴木　優† 吉野幸一郎† 中村　哲†

† 奈良先端科学技術大学院大学情報科学研究科　〒 630–0192 奈良県生駒市高山町 8916-5
E-mail: †{ysuzuki,koichiro,s-nakamura}@is.naist.jp

あらまし　Wikipedia の質を測定するとき，編集者の履歴を用いる手法を考える．編集者の編集には，異なる意味に書き換える場合や同じ意味で文言修正を行う場合があるが，これらを機械的に判定することは難しい．そこでクラウドソーシングを用いることにより，人手による判定を行う．本提案により，編集者の質測定精度が向上する．
キーワード　Wikipedia，質，編集履歴，クラウドソーシング

Assessment of Editor Quality of Wikipedia using Crowdsourcing

Yu SUZUKI†, Koichiro YOSHINO†, and Satoshi NAKAMURA†

† Graduate School of Information Science, Nara Institute of Science and Technology　Takayama 8916–5, Ikoma, Nara, 630–0192 Japan
E-mail: †{ysuzuki,koichiro,s-nakamura}@is.naist.jp

Abstract　In this paper, we propose a method for measuring quality of Wikipedia articles using edit history. When Wikipedia editors edit texts, the editors do not always change grammatical errors or term usages but also change the meaning of the texts. However, we can not detect which type of changes automatically. We propose a method of using crowdsourcing techniques for detecting types of changes. Using our proposed method, we can improve the accuracy of measuring quality of articles.
Key words　Wikipedia, Quality, Edit History, Crowdsourcing

1. はじめに

Wikipedia などの共同執筆コンテンツでは，その編集者が悪意を持っている場合や十分な知識が無い場合に，質が低いテキストが記述される場合がある．ところが，それらのテキストは質が高いかどうかを判定することは，そのテキストに関する知識が必要な場合があることや，テキストの投稿が大量であることなどから困難な作業である．そのため，そのような質が低いテキストを自動的に抽出することは，有用であると考えられる．そこで本研究では，自動的にテキストが不適切であるかどうか判定することを目的とする．

本研究ではテキストの質を，そのテキストが Wikipedia 上に存在することがふさわしいと考える人の割合であると定義する．つまり，テキストの質は内容の真偽に依存しない．あるテキストが多くの人によって正しいと信じられており，かつ日本語の完成度として優れているとき，そのテキストは質が高いとする．

Wikipedia のテキストから適切であるかどうかを自動的に抽出するための方法として，編集履歴を用いた方法が数多く提案されている．この手法では，Wikipedia では不適切な編集がなされたとき，他の編集者によってそのテキストが削除されたり変更されるという性質を利用している．編集者の質が高ければその編集者のテキストは残存することが多く，編集者の質が低ければその編集者のテキストは削除されることが多い．つまり，編集者の質が高いかどうかを，テキストの追加・削除だけで判定することができる．

これらの手法では，多くの編集者がテキストを変更しなければ，そのテキストの質が高いと考えられる．ところが，テキストが変更されたときであっても必ずしも意味内容が変更されているわけではない．そのためテキストの質を測定するためには，テキストの変更による意味内容の変化を観察する必要がある．ところが，テキストの意味が変更されているかどうかを自動的に判定することは困難である．一方で，人間がテキストの変更を観察する場合，その人が持つ知識を利用することによって意味が変更されているかどうかを容易に識別することができる可能性がある．

そこで本研究ではクラウドソーシングを用いることによって，人手によってテキストの意味が変更されているかどうかを識別し，テキストが不適切であるかどうかを判定する手法を提案する．提案手法を用いることによって，自動的に質の測定を行うよりも質の算出精度が向上すると考えられる．

クラウドソーシングとは，ワーカと呼ばれる不特定多数の作業者に対して小規模の作業を依頼する仕組みである．作業の依

頼者は，大規模な作業を低コストで短期間のうちに終わらせることができる利点があり，ワーカは空いた時間で短時間の作業を行い収入を得ることができる利点がある．そのため，機械学習における学習データ構築や評価実験における正解データ作成など，計算機では正確な処理を行うことが困難であるが人間にとっては容易な，大量の作業量が必要な作業を実現するために利用されている．

クラウドソーシングでは，大量のワーカが作業を行うため高速に処理されるという利点がある一方で，正確な作業を行わないワーカにより作業の精度が低下するという問題がある．これらの問題を解決する方法として，正解集合を用いる方法と用いない手法が従来手法として提案されている．正解集合を用いる方法では，あらかじめ作業内容と作業結果を複数個準備しておき，ワーカが準備された作業結果通りの作業を行うかどうかでワーカの作業の正確性を測定する方法である．この手法は，あらかじめ作業結果を準備する必要があるため手間がかかることや，作業結果が準備されていない作業内容をワーカが行うとき，作業の正確性を測定することができないという問題がある．そこで，本研究では正解集合を用いずにワーカの作業精度を測定する方法を用いた．具体的には，同一の作業内容に対して複数の作業者が同一の作業結果を行ったかどうかを測定した．

2. 関 連 研 究

Wikipedia では，様々な指標を用いて記事の評価が行われている [1] が，編集者に対して直接評価を行うものは少ない．そこで，編集者に対しての評価を行う方法を述べたあと，記事の評価に関する手法について述べる．

研究者を評価する指標の一つに，研究者を論文の質によって評価する指標がある．たとえば有名なものとしてインパクトファクター [2] がある．この指標は，学術雑誌に対して質が高いかどうかを計測するための方法であり，多く引用される論文は良いという仮定に基づいている．つまり，引用という論文著者の活動から，評価を推測していることになる．Wikipedia においても同様に，この考え方を基に編集者の質を測定するための主要な手法として，相互評価 (Peer review) を用いる手法が提案されている．この手法は，編集者相互や記事相互で評価し，質の算出を行う方法であるが，Wikipedia では編集者が直接他の編集者に対して評価を行わないため，編集者の編集活動から編集者間の評価を推測する必要がある．このような間接的な評価は Wikipedia 以外にも様々な分野で行われてきた課題であり，様々な解決方法が提案されてきた．主要な手法として，記事間のリンク関係を用いる手法，およびテキストの残存を利用した手法が挙げられる．

本研究と同様に，編集者間の相互評価による記事の質の算出手法も提案されている．Adler ら [3]〜[5]，Hu ら [6] は，編集履歴を用いて質の算出を行う方法を提案している．Wikipedia において記述が行われたとき，その記述の質が高いときには他の著者から削除されないが，質が低いときには削除されることが多い．この仮定は，West ら [7] による編集履歴の解析結果からも，成立する場合が多いことが期待できる．つまり，多くの

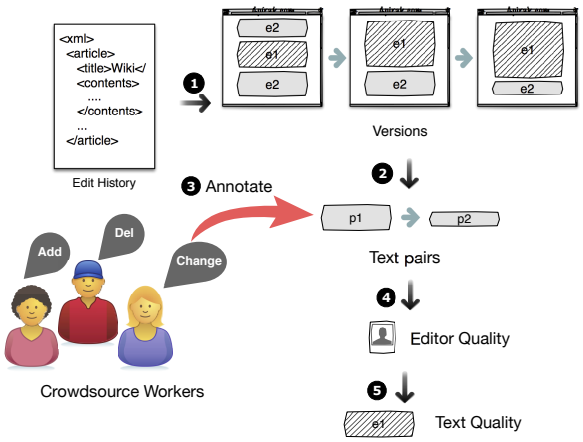

図 1　システムの全体図

編集を経て残存する記述は質が高いと推定することができる．そこで Adler らは，編集履歴を用いて記述の編集における残存率を計算することによって，質の計算を行っている．我々も，この仮定に基づいて記事の質測定を行った [8]〜[11] が，この際の問題点として，偶発的に残存するテキストによって編集者の質が大幅に変化することが挙げられる．提案手法では，このような偶発的な事象によらずに編集者の質を算出することができるため，精度の高い編集者の質を求めることができると考えられる．

3. 提 案 手 法

本研究で提案するシステムの全体図を，図 1 に示す．本システムでは，以下のような手順でテキストの質を算出する．

（1） 編集履歴からテキストペアを抽出
（2） テキストペアの変更種別を判定
（3） テキストペアの変更種別から，編集者が他の編集者へ行う編集意図を推定
（4） 編集者の質を推定
（5） テキストの質を算出

以下，それぞれの手順について以下で述べる．

3.1 テキストペア抽出

編集履歴から，編集により意味内容が変化している可能性があるテキストペア集合 $P = \{p_1, p_2, \cdots, p_N\}$ を抽出する．

まず，記事 d の編集履歴 $V = \{v_0, v_1, v_2, \cdots v_n\}$ を編集履歴ファイルから抽出する．v_i は i 番目に行われた編集の結果となるテキストであり，編集された時刻の早いものから順に時系列で並べられている．v_0 は空のテキストである．これらの編集履歴から，編集前と編集後のテキストペア群を抽出する．ここで抽出されるテキストペアとは，編集前と編集後で同じ文章が変更されたと考えられるテキストのペアを指す．そのため，編集によって一つの文章が二つ以上に分割された場合や，二つ以上の文章が一つに統合された場合などは，本研究ではテキストペアに含まない．

ここでは任意のバージョン v_i と v_j $(i, j \in \{0, 1, \cdots, n\}, i < j \leqq i + \alpha)$ からテキストのペアを取り出すことを考える．ここで α はパラメータであり，あるバージョンからどれくらい後の

バージョンまで比較するかを示す値である．本研究では $\alpha = 5$ と設定した．つまり，対象バージョンの次バージョンから 5 バージョン先までの 4 バージョンを比較対象とした．

まず，v_i を文章単位に区切る．本研究では，句点と記号を区切り文字にすることにより，テキストを文章単位に区切った．区切られた結果となる文章群を $S_i = \{s_{i,1}, s_{i,2}, \cdots, s_{i,m_i}\}$ とする．同様の手法を用いて，v_j から文章群 S_j を作成する．

最後に，S_i と S_j に含まれる文章からテキストペアを抽出する．まず，S_i から一つの文章 $s_{i,k}$ を取り出し，S_j に含まれるどの文章と対応するかを考える．まず，$s_{i,k}$ と S_j に含まれる全ての文章との類似度を算出する．本研究では，それぞれの文章を形態素単位に分割し，名詞のみを取り出し，出現頻度 (Term Frequencey; TF) に基づき出現頻度ベクトルを文章ごとに算出する．そして，ベクトル間の類似度をコサイン類似度により算出する．このとき，$s_{i,k}$ に最も類似している S_j 内の文章が $s_{j,l}$ であり，$s_{i,k}$ と $s_{j,l}$ との類似度が $sim(s_{i,k}, s_{j,l})$ であったとする．このとき，$sim(s_{i,k}, s_{j,l})$ があるパラメータ sim_{min} 以上 sim_{max} 未満である場合に，文章間に対応関係が存在すると考えられるためテキストのペア $s_{i,k}$， $s_{j,l}$ を集合 P に加える．このとき，文章間の類似度が小さすぎる場合には，異なる文章間でペアを作成してしまう場合があり，抽出するテキストのペアとして不適切である．一方，文章間の類似度が大きすぎる場合には，ほぼ同じ文章であるため変更点が無く，やはり抽出したいテキストペアとして不適切である．以上の操作を，S_i に含まれる全ての文章に対して行う．

3.2 テキストペアの変更種別判定

3.2.1 節で算出したテキストペア集合 P に対して，変更種別を付与する．まず，テキストペアに対して付与する種別について述べる．次に，クラウドソーシングによらずに変更種別を付与する方法について述べる．最後に，クラウドソーシングを用いた変更種別の付与手法について述べる．

3.2.1 変 更 種 別

本研究では，クラウドソーシングにより文章の意味が変更されたかどうかだけに限定してワーカによる判定を行った．つまり，ワーカは文章の質が改善もしくは改悪されたか，正しいことが書かれているか誤っているかなどの判定を行っていない．なぜならば， 1. 章で述べたように Wikipedia のテキストは質が高いかどうかを判定するためには，そのテキストに関する知識が必要となる場合が多いためである．

本研究で用いるテキストの変更種別とは，以下の四種類であると定義する．

- 内容を追加・詳細化
- 内容を削除・簡略化
- 内容を変更し異なる意味へ変化
- 内容変更したが意味は不変（誤字の修正など）

ここで，変更種別はテキスト意味内容だけに着目し，日本語としての表現の変化を排除した．例えばあるテキスト上で「為に」が「ために」に変更されたとき，漢字からひらがなに変更されたが意味内容は変化していないため，不変であると判定する．

3.2.2 クラウドソーシングによらない変更種別特定

3.2.1 節で算出したテキストペア集合 P に含まれるテキストペアの中には，自動的に変更種別を特定できるものが二つある．一つは編集前と編集後で形態素が単に増加した場合，もう一つは単に減少した場合である．これらに該当するテキストペアに対しては，それぞれ内容の追加，内容の削除であると考え，変更種別を付与する．

ところが，変更後に追加も削除も行われた場合には，変更種別を自動的に付与することが難しい．そこで，クラウドソーシングにより変更種別を付与する．

3.2.3 クラウドソーシングを用いた変更種別判定

テキストペア集合 P のうち，3.2.2 節で特定されなかったテキストペアに対して，クラウドソーシングにより人手で変更種別を判定する．手順は次の通りである．

（1） ワーカにテキストペアを提示
（2） ワーカによりテキストペアに対して変更種別を指定
（3） ワーカが付与した変更種別を集約

まず，システムはワーカに対してテキストペアとして，編集前と編集後の二つのテキストを提示する．ワーカが付与する質問は以下 Q1 と Q2 の二つである．

（1） Q1: 文から得られる情報・内容はどのように変化しましたか?

- 追加し，詳細になった．
- 削除し，簡略になった．
- 変更し，異なる意味になった．
- 変更したが，意味はあまり変わっていない．(誤字の修正を含む)
- 対応していない．違う文章の組だと思われる．
- 変更前もしく変更後の文章は意味が分からない．

（2） Q2: 変更前の文章と変更後の文章では日本語としての表現は変化（改善もしくは改悪）しましたか?

- 分かりやすくなった．もしくは誤字が修正された．
- 変わらない．
- 分かりにくくなった．もしくは変更後の文が誤りを含むように変わった．
- 変更が大きい，対応していない，意味が分からないなどの理由で判定できない．

ここで Q1 はテキストの意味内容の変更についての設問であり，Q2 は日本語の文法や語彙についての完成度についての設問である．まず，Q1 と Q2 の二つを分けて設問として設定した意図を述べる．あるテキストが変更されたとき，そのテキストの変更は意味内容を含む場合，表現だけの修正に留まっている場合，その両方を含む場合の三つに分類することができる．本研究ではまず意味内容の変更だけを考慮するため，表現の修正については別の設問とすることにより意味内容の変更だけを判定した．

このとき，一つのテキストペアにつき少なくとも二人以上のワーカによって変更種別を付与する．二人の判定が Q1 において同一であった場合，その判定を採用するが，異なる場合にはもう一人のワーカを追加して，判定を行った．このとき，三人

のワーカのうち二人が同じ判定を行った場合，多くのワーカによる判定を採用する．三人のワーカが全て異なる判定を行った場合，さらに一人のワーカを追加して判定を行った．以上のような手順を用いてワーカを追加し，過半数のワーカが同じ判定を行った場合にはその判定を採用した．五人以上のワーカを追加しても判定が割れ，50% 以上のワーカが同じ判定を行った場合には，そのテキストペアの判定そのものが困難であると判定し，判定対象から外した．

Q1 において，3.2.1 節で設定した選択肢以外に，対応していない場合と意味が不明である場合の二つの選択肢を追加した．なぜならば，本システムにおいてテキストペアの精度が十分ではない場合が存在するためである．本システムでは，単純な方法として句点や記号による区切りを文の区切りとして利用した．そのため，そもそも文として成立しないテキストペアも存在した．特に，Wikipedia において特有の記号列などについては，それらの対応を行わなかった結果，意味が不明であるテキストペアも存在している．それらは人間によって判定することが容易であるため，ワーカにより判定を行った．

以上の手順を用いて，テキストペア p に対して最も多くのワーカが付与した判定を $t(p)$ として付与する．

3.2.4 クラウドソーシングにおける判定の正確性向上

クラウドソーシングにおいてワーカによる判定精度を向上させるため，以下の三つの方法を用いた．

（1） 例となるテキストペアと判定結果の例示
（2） 他のワーカによる判定の表示
（3） 他のワーカとの判定の一致度合いの表示

一つ目の判定結果の例示として，全てのワーカに対して判定基準を伝えるために，あらかじめ七つのテキストペア，およびその判定結果を，作業を始める前に閲覧させた．これら七つのテキストペアには，明白に判定を得られるものと，不明確であり基準が必要となるものを選んだ．ところが，基準を明確に明文化してワーカに示す方法は利用しなかった．なぜなら，テキストペアには様々な分類があり，全ての場合における基準を明文化することは難しいと考えられたためである．また明文化しないことにより他のワーカが協調し，多くのワーカが一般的にどのように考えるかという基準を，自動的に構築することができるかどうかを試したためである．

そこで，ワーカ自身の判定が他のワーカの判定と基準が似ているかどうかをワーカ自身で判定するために，判定後に他のワーカがどのような判定を行っているかという情報を提示する．ところが，他のワーカがまだ判定していないテキストペアをあるワーカが判定したとき，その判定が他のワーカの判定と同じかどうかが分からない．そこで，他のワーカが判定した後で改めて他のワーカの判定と一致しているかどうかを確かめることができるように，システムを構築した．

3.3 テキストの質算出

3.2.2 節および 3.2.3 節で求めたテキストペアへの判定を利用して，テキストの質を算出する． 1. 章で述べた通り，編集者の質が高いときその編集者が記述したテキストの質は高いと考える．そこで，Adler らの手法 [3] や著者ら [8] における手法に編集者の準じた手法を用いて編集者の質を算出する．ある編集者 e が Wikipedia の記事 d における質 $q(e,d)$ を以下のように定義する．

$$q(e,d) = \frac{|r(e,d)|}{|s(e,d)|} \tag{1}$$

ここで，$r(e,d)$ は，記事 d において編集者 e が他の編集者によって意味内容を変更された文章集合であり，3.2.3 節において「変更し，異なる意味になった.」と一度以上判定された文章の集合である．また，$s(e,d)$ は編集者 e が記述した全文章の集合である．$|r(e,d)|$，$|s(e,d)|$ はそれぞれ文章集合 $r(e,d)$，$s(e,d)$ に含まれる文章の数である．

編集者 e の質 $Q(e)$ を以下のように定義する．

$$Q(e) = \frac{\sum_{d \in D(e)} q(e,d)}{|D(e)|} \tag{2}$$

ここで $D(e)$ は編集者 e が Wikipedia に記述した記事集合であり，$|D(e)|$ は $D(e)$ に含まれる記事数である．

以上の手順により，編集者に対して質を求める．最後に，テキストに対してその編集を行った編集者を特定し，テキストの質はそのテキストを追加した編集者の質であると定義する．

4. 予 備 実 験

現在，テキストの質を求めるために，評価実験を行っている．評価実験には二つあり，一つはクラウドソーシングにより適切な判定を行うことができるか，もう一つは適切にテキストの質を判定することが可能であるかどうかを測定する予定である．本稿ではそのための予備実験として，クラウドソーシングによりどのような判定が行われているかを示す．

4.1 評 価 手 法

評価実験には，2015 年 5 月 12 日時点での Wikipedia 日本語版の編集履歴データ (注1) を利用する．Wikipedia における記事の記述はその版ごとに，記事の全記述および編集された日時や編集者が記録され，編集履歴データとして定期的に Web 上に公開されている．

これらの編集履歴から，判定対象となるテキストペアを抽出した．編集履歴に含まれる記事の数は 1,523,561 件であり，編集者数は 3,016,675 人である．このうち，50 回以上の編集が行われている記事数は 192,812 件であったため，これらの記事からテキストペアを抽出した．その結果，削除と追加を両方含むテキストペアを 1,060,090 個抽出した．これらのテキストペアは自動的に判定することが困難であるため，クラウドソーシングにより判定を行った．

クラウドソーシングでは，クラウドワークス (注2) において募集したワーカにより作業を行った．1ヶ月程度の募集期間で，699 名のワーカにより 19,409 個のテキストペアに対して 50,836 件

(注1)：http://dumps.wikimedia.org/jawiki/20150512/
(注2)：http://crowdworks.jp

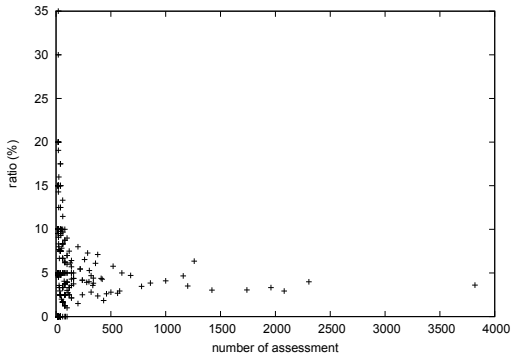

図 2　ワーカ一人の評価回数とそのワーカが判定したテキストペアが「不一致」となった割合の関係

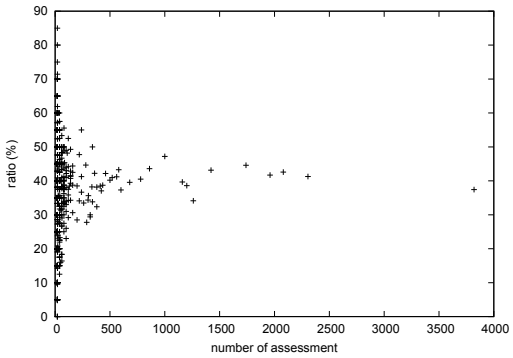

図 3　ワーカ一人の評価回数とそのワーカが判定したテキストペアが「一致」となった割合の関係

の評価を行った．このとき，人手により作業内容を確認し，不正確な評価を行っているワーカに対しては報酬を支払わなかった．このときの判定基準として，作業時間や他のワーカによる判定結果の一致度，確認を行った作業者による判定などを総合的に判断した．

4.2 評価結果

表 1 において，テキストペアごとにワーカの評価がどの程度一致，不一致しているかを示す.「一致」は，テキストペアを評価した全てのワーカが同じ評価を行ったテキストペアの数,「準一致」は，テキストペアを評価したワーカのうち半分以上が同じ評価を行ったテキストペアの数，そして「不一致」は全てのワーカが異なる評価を行ったテキストペアの数を示す．なお 3.2.3 節で述べたように，テキストペアは最低 2 名以上のワーカに割り当てられ，一つの変更種別を選択したワーカが過半数を超えた場合か，ワーカの数が 4 名以上となるまで，ワーカを新たに割り当てる．そのため，上記の分類以外の変更種別の分布となることは無い．

この表から，97.12% のテキストペアに対して過半数のワーカから同一の評価が行われており，ほとんどのテキストペアに対してクラウドソーシングにより変更種別を特定可能であることが分かった．

4.2.1 ワーカごとの判定分析

変更種別の判定がワーカにより一致しない原因として，以下の二つが考えられる．

（1） 一部のワーカが多くのワーカと異なる判定を行う

（2） 判定する問題が難しい

まず，(1) に相当するワーカがどの程度存在するかを確かめるため，ワーカごとの「不一致」と「一致」の割合をそれぞれ図 2, 3 に示す．これらの散布図では，一つの点は一人のワーカを指している．この図からは外れ値がほとんど観測されないことからも，本実験では全体のワーカと異なる判定を行う一部のワーカは存在しないと考えられる．

表 1　ワーカによる評価の一致傾向

分類	判定数	割合
一致	9,962	51.33%
準一致	8,887	45.79%
不一致	560	2.89%
合計	19,409	100%

(2) のような場合が存在するかどうかは,「不一致」に分類された設問を精査する必要がある．そこで，設問とその判定結果について述べる．

4.2.2 ワーカによる判定の不一致についての考察

判定を行うことが難しいテキストペアは，以下の二つに分類できる．

- テキストペアが正確に抽出できていない．
- 複数種類の変更点が存在しており，最も重要と考える変更点がどの点であるかがワーカにより異なる．

一つ目の問題は，Wikipedia 記法によって記述された図表部分など記号が数多く存在する部分から，誤って文とはならない部分を文として抽出してしまうということが一つの原因である．もう一つの原因として，テキストペアのうち変更前と変更後で意味内容が全く異なる場合があることが挙げられる．これらのテキストペアは,「対応していない」や「意味がわからない」などの変更種別をワーカが選択すると考えていた．ところが，多くのワーカはそのような不完全な文とワーカ自身の知識を組み合わせ，変更点に何らかの意味を見い出そうとした形跡が見られる．ワーカが持っている知識はそれぞれ異なり，しかも不完全な文から推測される文はワーカごとに異なるため，変更種別がそれぞれ異なった．

二つ目の問題は，以下のような事例で説明する．

記事: 日本語

変更前　中華人民共和国中国の現代漢語常用字表では、「常用字」として 2500 字、「次常用字」として 1000 字が定められており、日中で日常使われる漢字の字数にそれほど大きな隔たりはないといえる。

変更後　なお、漢字の本家である中華人民共和国においても同様の基準は存在し、現代漢語常用字表により、「常用字」として 2500 字、「次常用字」として 1000 字が定められている。

このテキストペアは四人のワーカにより評価が行われた．この例では，二つの文間の差分が大きく，常用字として 2500 字，次

常用字として 1000 字が定められている，という点以外は異なる文章である．このとき，あまりにもテキストの差分が多いことから，テキスト同士がそもそも対応しておらず，異なる主旨のテキストであると判定しているワーカが存在する．また，変更前にあった日本や中国で日常使われる漢字の字数に大きな差がないという部分が，変更後には削除されていることから，内容を削除，簡略化したと判定したワーカが存在した．さらに，差分に対して意味的な差を感じなかったワーカは「変更したが，意味はあまり変わっていない.」と判定しており，意味的な差を感じたワーカは「変更し，異なる意味になった.」を選択している．

このテキストペアにおける例の通り，文の内容が複数箇所変更されたとき，どのような評価を行うかはワーカや Wikipedia 読者の主観によるところが大きい．具体的には，追加された文と削除された文，および共通している文に対してどのような重要度を付与するかという点が，ワーカの判定に大きく依存していることが分かった．

4.2.3 不適切なワーカの判別

本実験結果から，ワーカごとのテキストペアに対する判定の「一致」率や「不一致」率を用いて，不適切なワーカであるかどうかを判定することが可能となると考えられる．図 2，3 から，ワーカ一人あたりの評価回数が増えるごとに，そのワーカによる判定の「不一致」率は 3.5%,「一致」率は 42% に収束することをが分かる．また，ワーカが 500 回程度の評価を行うと，ほぼこれら二つの割合は収束値から大きく外れた値にはならないことが分かる．つまり，これらの割合と収束値を比較することによって，ワーカが不適切かどうかを客観的に判定することができる．

ところが，ワーカの評価回数が 500 回以下となる場合,「一致」や「不一致」の割合はワーカによる判定の適切度と相関関係が無い．クラウドソーシングでは，評価回数に応じて作業料をワーカへ支払うため，早期にワーカの判定が適切であるかどうかを判定する必要があるが，今後の課題である．

5. おわりに

本研究ではクラウドソーシングを用いることによって，高精度に Wikipedia におけるテキストの質を求めるための手法を提案した．テキストの質を求める際には，編集者が編集元のテキストに対して残存させようとしているのか，意味的に変更しようとしているのかを判定することが重要である．ところが，テキストに含まれる形態素の変更だけではなく，その形態素が持つ意味や文の意味，その文が持つ背景知識などを持っていなければ，これらの判定を行うことは難しい．そこで本研究では，クラウドソーシングにより，文の変更が行われたときの意味内容の変化について，判定を行った．

予備実験の結果，クラウドソーシングにより適切な判定を行うことができる可能性を示唆することができた．その一方で，テキストの意味的な変更とは主観的なものであり，ワーカが持つ背景知識や読解能力などに依存する判定も存在することが分かった．

今後の課題について述べる．クラウドソーシングによる判定には，金銭的なコストがかかること，作業速度の上限には大きな限界があることが問題点である．今回の場合，比較すべきテキストペアの数は 1,060,090 件であり，一件 1 円で処理を依頼したとしても 300 万円以上と極めて大きな金額が必要となってしまう．そこで，現在までに得られた評価を教師とした教師あり学習手法を用い，機械学習により自動的に判定を行う方法が考えられる．

さらに，クラウドソーシングにおける問題点である，判定精度が低いワーカを自動的に発見する手法を考案する必要がある．現在は人手により発見を行っているが，作業コストが高いため，さらに大規模なクラウドソーシングを行った場合には適用が難しくなる可能性がある．他のワーカによる判定とどの程度一致した判定を行っているか，どれだけの時間をかけて判定を行っているかなど，判定に付随する様々なメタデータを利用することによって，ワーカの質を判定することができるのではないかと考えている．

謝　　辞

本研究の一部は，NAIST 多元ビッグデータ解析による知の創出研究拠点事業，および JSPS 科研費（23700113）による．

文　　献

[1] 鈴木優. Wikipedia における情報の質. 情報処理学会論文誌: データベース, Vol. 6, No. 4 (TOD56), pp. 46–58, 2013.

[2] E Garfield. Citation analysis as a tool in journal evaluation. *Science*, Vol. 178, No. 60, pp. 471–479, 1972.

[3] B.T. Adler and L. de Alfaro. A content-driven reputation system for the Wikipedia. In *Proceedings of the 16th international conference on World Wide Web (WWW '07)*, pp. 261–270, 2007.

[4] B. Thomas Adler, Krishnendu Chatterjee, Luca de Alfaro, Marco Faella, Ian Pye, and Vishwanath Raman. Assigning trust to wikipedia content. In *Proceedings of the 4th International Symposium on Wikis*, WikiSym '08, pp. 26:1–26:12, New York, NY, USA, 2008. ACM.

[5] B. Thomas Adler, Luca de Alfaro, Ian Pye, and Vishwanath Raman. Measuring author contributions to the wikipedia. In *Proceedings of the 4th International Symposium on Wikis*, WikiSym '08, pp. 15:1–15:10, New York, NY, USA, 2008. ACM.

[6] M. Hu, E. Lim, A. Sun, H. W. Lauw, and B. Vuong. Measuring Article Quality in Wikipedia: Models and Evaluation. In *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM 2007)*, pp. 243–252, 2007.

[7] Andrew G. West and Insup Lee. What wikipedia deletes: characterizing dangerous collaborative content. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, pp. 25–28, New York, NY, USA, 2011. ACM.

[8] Yu Suzuki. Effects of implicit positive ratings for quality assessment of wikipedia articles. *Journal of Information Processing*, Vol. 21, No. 2, pp. 342 – 348, 2013.

[9] Yu Suzuki and Masatoshi Yoshikawa. Mutual evaluation of editors and texts for assessing quality of wikipedia articles. In *Proceedings of the International Symposium on Wikis and Open Collaboration (WikiSym 2012)*, p. online, 2012.

[10] Yu Suzuki. Assessing quality values of wikipedia articles using implicit positive and negative ratings. In *Proceedings of the 13th International conference on Web-Age Information Management (WAIM 2012)*, pp. 127 – 138, 2012.

[11] Yu Suzuki and Masatoshi Yoshikawa. Qualityrank: Assessing quality of wikipedia articles by mutually evaluating editors and text. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media (HT 2012)*, pp. 307 – 308. ACM Press, 2012.