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
IN2015-98 (2016-01)

文章校正における共起語を用いた漢字の誤変換の検出

梶谷　貴士†　　服部　峻††

†,†† 室蘭工業大学 ウェブ知能時空間研究室　〒 050–8585 北海道室蘭市水元町 27–1
E-mail: †12024035@mmm.muroran-it.ac.jp, ††hattori@csse.muroran-it.ac.jp

あらまし　既存の文章校正ツールによる文章中の漢字の誤変換の指摘は，予め用意された誤変換の用例と合致するか否かで判断しているものが多い．しかしながら，このような方法では，予め用意された誤変換の用例集に含まれない未知の漢字の誤変換を指摘することはできない．そこで本稿では，入力された文を形態素解析して切り出した文節ごとに変換候補を求め，各文節に対する複数の候補の中から，その文節の近傍に存在している文節群との共起性が最も高いものを選択することによって，その文章の文脈に相応しい，正しい変換を精確に導き出すシステムを提案する．文節同士の共起性の指標である共起度は，日々増大して行く Web 上のページ群を活用して算定する．また，提案システムは，多くの既存の文章校正ツールとは異なり，予め用意された誤変換の用例を使わないため，未知の漢字の誤変換に対しても検出できる可能性がある．評価実験として，文中に漢字の誤変換を必ず 1 つのみ含む文 100 個とその誤変換を正しく変換した同じ文 100 個を用意し，計 200 個の文を提案システムに入力して，漢字の誤変換の検出精度を測定した．その結果，パラメータに依って最大で 62%という誤字訂正率と，一様に 4%という正字誤訂正率が得られた．
キーワード　誤変換検出, 共起性, 文章校正, 形態素解析, Web マイニング

Detection of Mis-converted Chinese Characters in Text Proofreading by Co-occurrence Words

Takashi KAJIYA† and Shun HATTORI††

†,†† Web Intelligence Time-Space (WITS) Laboratory, Muroran Institute of Technology
27–1 Mizumoto-cho, Muroran, Hokkaido, 050–8585, Japan
E-mail: †12024035@mmm.muroran-it.ac.jp, ††hattori@csse.muroran-it.ac.jp

Abstract　Most of the existing tools for text proofreading detect mis-converted Chinese characters in a target text by judging based on whether or not they match the prepared example(s) of mis-conversion. However, such a method cannot detect unknown mis-converted Chinese characters that do not exist in the prepared examples of mis-conversion. Therefore, this paper proposes a novel system that extracts clauses by morphological-analyzing an input sentence, and acquires a contextualized conversion for each clause by choosing its one candidate which have the greatest co-occurrence with the surrounding clauses. The proposed system assesses the degree of the co-occurrence between clauses by using enormous pages in the exponentially-growing Web. And the system has the capability of detecting unknown mis-converted Chinese characters, because it does not have to prepare a set of examples of mis-conversion unlike most of the existing tools for text proofreading. By inputting 200 sentences of 100 sentences with only one mis-conversion and the corrected 100 sentences without the one mis-conversion to the proposed system, the evaluation experiment measures its precision of detecting mis-converted Chinese characters. As a result, the system has achieved 62% at the most for the ratio of true alarm, and 4% stably for the ratio of false alarm, depending on its parameter of the number of Web pages for assessing the co-occurrence.
Key words　Mis-conversion Detection, Co-occurrence, Text Proofreading, Morphological Analysis, Web Mining

1. まえがき

現在，インターネット上には数多くの文章校正ツール [1–3] が存在しているが，漢字の誤変換を指摘する機能が含まれているものは極僅かしかなく，誤変換を検出する方法も予め用意された誤変換の用例と比較して，入力された文章の中に誤変換の用例と同じ部分が含まれているかどうかで判定している．例えば，入力された文章の中に「以外と人数が揃わない」という漢字の誤変換が含まれていた場合，

「以外と人数」 $\Longrightarrow$ 「意外と人数」

という誤変換の用例とその訂正例が文章校正ツールに登録されていれば検出して指摘することができる．しかしながら，この用例だけでは「以外と知られていない」という漢字の誤変換には対応することができないため，

「以外と知られていない」 $\Longrightarrow$ 「意外と知られていない」

という用例を新たに追加登録しなければならない．一方，これらの問題を一括でまとめて回避すべく，

「以外と」 $\Longrightarrow$ 「意外と」

という誤変換の用例とその訂正例をもし文章校正ツールに登録してしまうと，「彼以外と映画に行くのは」という正しい変換に対してまで間違いとして指摘してしまう．従って，後者のようにマッチング条件として緩い漢字の誤変換の用例とその訂正例で一括して対応するのではなく，前者のように，より具体的な漢字の誤変換の用例とその訂正例を出来る限りたくさん想定して文章校正ツールに予め登録しておかなければならず，データ量が膨大になるだけでなく，新しく生まれた未登録の誤変換の用例も追加して行くという随時メンテナンスも行わなければならないため，より柔軟な誤変換検出手法が必要である．

そこで本稿では，事前にシステムに登録した漢字の誤変換の用例と比較するのではなく，入力された文を形態素解析して切り出した文節ごとに変換候補を求め，各文節に対する複数の候補の中から，その文節の近傍に存在している文節群との共起性が最も高いものを選択することによって，その文章の文脈を考慮した正しい変換を精確に導き出すシステムを提案する．ある文節に対して共起性の高い語（共起語）を見つけ出すため，日々増大して行く Web 上にある膨大な文書群を参考にすることで，文節同士の共起度を算出する．

2. 提案手法

本章で詳述する提案システムは，まず，ユーザからシステムに入力された文章（本稿の実験では文）を形態素解析によって文節に切り分け，かな文字に開いた上で，文節ごとの変換候補を割り出す．次に，その変換候補の一つ一つと，文中における前後の文節との共起性をインターネット上の Web ページ群を参考にして調べ，文節ごとの複数の変換候補の中から共起性が最も高い変換候補をシステムの考える正解として選択する．最後に，文節ごとに正解として選択された変換候補のみを連結して文を再構成し，正解の文としてユーザに提示する．

以降，各手順について，順に詳しく述べて行く．

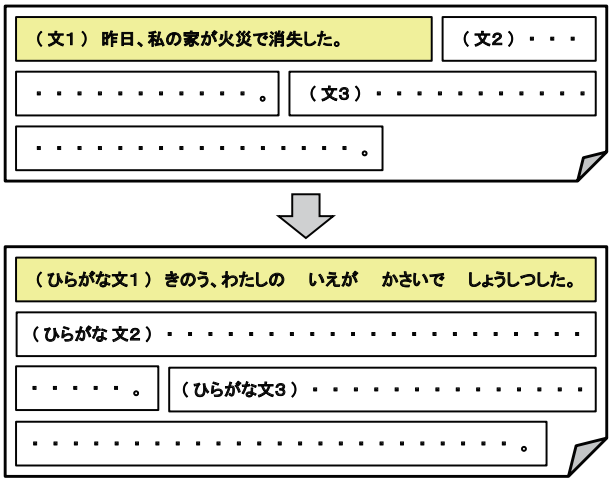

図 1　入力文をひらがな文に変換

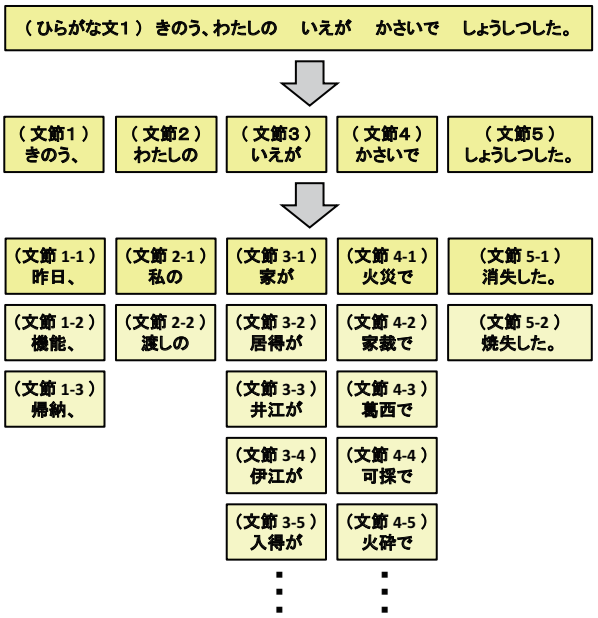

図 2　ひらがな文の各文節から変換候補をリストアップ

2.1　入力文の各文節における変換候補の割り出し

まず初めに，図 1 のように，ユーザから入力された，漢字の誤変換をチェックしたい文章（本稿の実験では文）を形態素解析 [4] することにより，全てひらがなで構成された文章に変換する．このとき，文節ごとに分かれるように，ひらがなだけで文を再構築する際，助詞と副詞，感嘆詞の直後に空白を入れておくようにする．この空白挿入の処理が必要な理由は，本稿の提案システムで使用している Yahoo!かな漢字変換 API [5] は，ひらがなで入力された文章を文節に分けて，文節ごとの変換候補を求めるものであるが，ひらがなのみで構成された文章では，ユーザからシステムに入力された元の文章の区切り方とは異なる区切り方で区切られてしまう可能性があり，このような問題が極力発生しないようにするためである．

次に，ひらがなのみになった文章を句読点で区切り，一文ずつに分ける．そして，図 2 のように，一文ずつ，Yahoo!かな漢字変換 API に掛けて，文節ごとの変換候補を求めて行く．なお，図 2 における「(文節 1-1)」「(文節 2-1)」「(文節 3-1)」

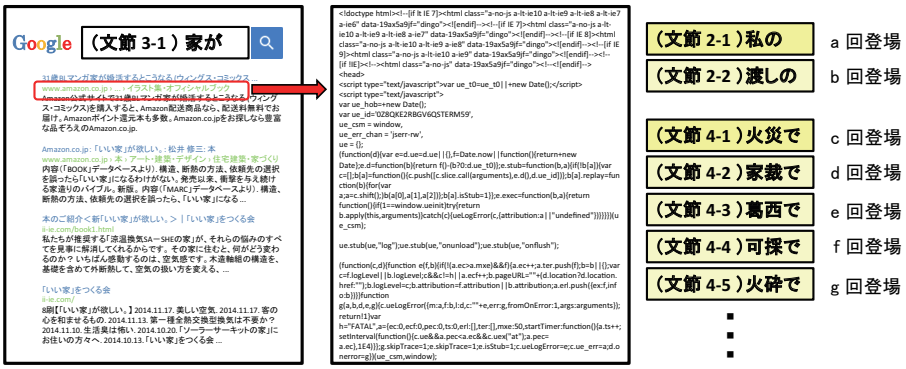

図 3　上位 N 件 Web ページ群における登場回数を用いた変換候補同士の共起性の算定

「（文節 4-1）」「（文節 5-1）」は，ユーザが提案システムに入力した際の変換と同じものが来るように設定されている．但し，ユーザがシステムに入力した際と同じ変換候補が Yahoo!かな漢字変換 API から返って来ない場合もあるが，その場合には，ユーザのシステムに入力した際の変換ではなく，Yahoo!かな漢字変換 API によって求められた一番上の変換候補が入る．

この時点で，変換候補のうち漢字が一切含まれていないものや，漢数字ではない数字が含まれているものは除外する．この処理を行う理由は，数字の部分がアラビア数字や漢数字，ローマ数字などが全て変換候補になってしまうのを避け，漢字の誤変換の検出が目的であるシステムへの負担を減らすためである．

さらに，文節全体が元々，ひらがなで入力されていたものであった場合には，ひらがなのみで構成された変換候補以外は全て除外され，ひらがなのみの文節のみが変換候補となる．例えば，指示代名詞の「その」は，「其の」「祖の」「租の」「園」「薗」「苑」「曽野」など変換候補が非常に多く，この処理を行わなかった場合，元々の「その」が共起性の高い変換候補として選択されることはまずない．また，提案システムとしても漢字の誤変換を検出することが目的であるため，元々ひらがなで入力されている部分は触れないようにするためでもある．

2.2　各変換候補と周辺の文節との共起性の算定

前節において，ユーザから入力された文を形態素解析して得たひらがな文，及び，そのひらがな文を Yahoo!かな漢字変換 API に掛けて求めた文節ごとの変換候補リストのセットに対して，順々に文節をずらして行く．入力文で i 番目の文節 c_i において，さらに順々に変換候補を切り替え，注目している変換候補 $c_{i,j}$ と，ユーザから入力された文における周辺（本稿では原則，前後のみ）の文節に対する各変換候補との共起性，例えば直前の文節に対する 1 番目の変換候補との共起性 $\mathrm{co}(c_{i,j}, c_{i-1,1})$ や，直後の文節に対する 1 番目の変換候補との共起性 $\mathrm{co}(c_{i,j}, c_{i+1,1})$ などを算定することで，注目している変換候補 $c_{i,j}$ の文脈としての影響を，周辺の文節に対する各変換候補 $c_{i-1,1}$ や $c_{i+1,1}$ などの相応しさの評価値に反映させて行く．

具体的には，例えば，図 2 の「（文節 3）いえが」に注目し，その文節の変換候補の一つである「（文節 3-1）家が」について，その変換候補が入力文における文脈として，周辺の文節に対する各変換候補にどれくらいの影響を与えるか，変換候補同士の共起性を算定する場合について取り上げる．まず，図 3 のように，変換候補「（文節 3-1）家が」を検索条件にして Google 検索を行い，その上位 N 件（本稿の実験では $N \in \{10, 20, 50, 100\}$）にランキングされた Web ページにアクセスして全体の文章を取得し，一時的に保存する．次に，その保存された Web ページの文章中において，注目している変換候補「（文節 3-1）家が」の前後の文節に対する各変換候補，例えば「（文節 2-1）私の」や「（文節 4-3）葛西で」などが，システムで設定された参照範囲内（本稿の実験では全文参照または一文参照）にいくつ含まれているかを数えて，前後の文節に対する各変換候補ごとに共起度 $\mathrm{co}(c_{3,1}, c_{2,1})$ や $\mathrm{co}(c_{3,1}, c_{4,3})$ として登場回数を記録する．

但し，文頭の文節 c_1 に対する変換候補の一つに注目している際は，直前の文節が存在しないため，直後の文節 c_2，及び，その次の文節 c_3 に対する各変換候補との共起性を算定する．同様の理由で，文末の文節に対する変換候補の一つに注目している際は，直後の文節が存在しないため，直前の文節，及び，さらに前の文節に対する各変換候補との共起性を算定する．

2.3　訂正文の作成と提示

全ての文節に対する各変換候補の検索，及び，上位 N 件にランキングされた Web ページ群を参照して求めた変換候補同士の共起性の算定が終了したら，ユーザから入力された文に対する各変換候補 $c_{i,j}$ の相応しさの評価値 $\mathrm{fitness}(c_{i,j})$ を以下の式に基づいて計算し，文節それぞれの変換候補のうち，Web ページに一番登場していた共起性の高い，言い換えると，入力文に対する相応しさの評価値が最も高い変換候補をシステムの考える正しい変換候補として採用する．

$$\mathrm{fitness}(c_{i,j}) = \sum_{k=1}^{n_{i-1}} \mathrm{co}(c_{i-1,k}, c_{i,j}) + \sum_{k=1}^{n_{i+1}} \mathrm{co}(c_{i+1,k}, c_{i,j})$$

ここで，n_i は，入力文で i 番目の文節に対する変換候補の総数を表している．例えば，図 2 の例では，$n_1 = 3$ である．

但し，変換候補全てヒットしていなかった場合，隣接する文節は共起性の無い文節同士と判断され，最初にユーザから入力された変換をそのまま採用する．そして，変換候補全てがヒットせず，かつ，最初にユーザが入力した変換も候補に無かった場合，Yahoo!かな漢字変換 API によって返された変換候補の中で一番最初に出て来たものを採用する．因みに，元々漢字が含まれていない変換候補で入力されていた文節については，事前に候補が一つになっているため，そのままの状態を正解とする．

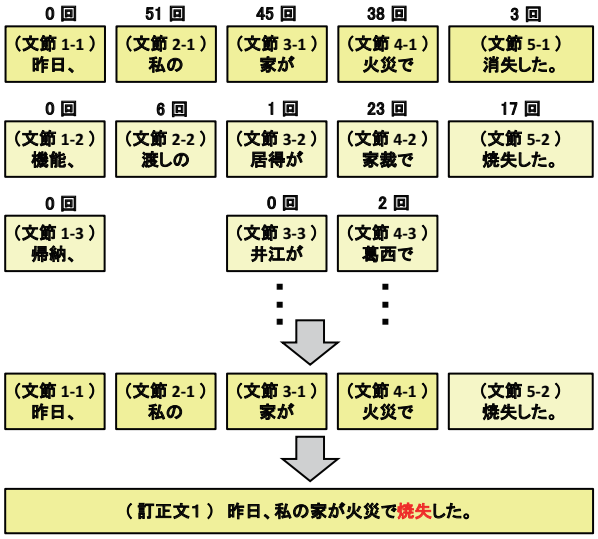

図 4　共起性に基づく相応しさの評価値を用いた訂正文の作成

最後に，図 4 のように，各文節に対して採用された最も相応しさの評価値が高い変換候補を全て繋ぎ合わせ，システムとしての訂正案として文章を作成し，ユーザに提示する．

3. 評 価 実 験

本章では，提案システムの漢字の誤変換の検出精度を評価するため，文中に漢字の誤変換を必ず 1 つのみ含む文を 100 文と，その 100 文の誤変換部分が本来の正しい変換に置き換えられている文との合計 200 文を入力し，漢字の誤変換を正しく訂正できた確率である誤字訂正率と，元々正しかった変換を誤って誤変換に訂正してしまった確率である正字誤訂正率とを求める．

変換候補同士の共起性を算定するステップで使用されるパラメータ N に関して，Google 検索における上位何件の Web ページを参照するのが良いか，上位 10 件，上位 20 件，上位 50 件，上位 100 件の 4 パターンで評価実験を行った．さらに，各 Web ページにおける参照範囲に関しても，全文章を参照するパターンである全文参照と，Google 検索した変換候補が含まれている一文のみを参照するパターンである一文参照の 2 パターンで，従って，計 8 パターンで誤字訂正率と正字誤訂正率を求めた．

表 1 や図 5 のように，全文参照と一文参照とを比較すると，変換候補同士の共起性の算定のために参照する上位 Web ページの数 N が同じ場合，全文参照の方が全ての場合において誤字訂正率が高くなった．また，同じ参照範囲を用いた場合，参照する上位 Web ページの数 N が多いほど誤字訂正率が高くなった．しかしながら，ほとんどの場合で半分以上の漢字の誤変換が検出できず，そのままスルーしてしまう結果となっている．

一方，正字誤訂正率に関しては，変換候補同士の共起性の算定のために参照する上位 Web ページの数 N や，参照パターンを変化させても，変わらない結果となった．これらには，誤訂正の原因が Yahoo!かな漢字変換 API が入力した際の文とは異なる部分で区切って作成してしまったために起きたケースが 2%，元は正しい変換であったにも関わらず間違いとして検出され誤った変換を提示してしまったケースが 2%あった．

表 1　提案システムの誤字訂正率と正字誤訂正率のパラメータ依存性

参照範囲	上位 N 件	誤字訂正率	正字誤訂正率
全文参照	10 件	31%	4%
	20 件	38%	4%
	50 件	51%	4%
	100 件	62%	4%
一文参照	10 件	20%	4%
	20 件	33%	4%
	50 件	43%	4%
	100 件	52%	4%

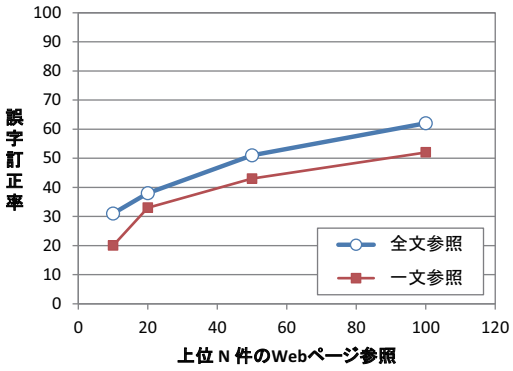

図 5　提案システムの誤字訂正率のパラメータ依存性

4. む　す　び

本稿では，入力文を形態素解析して切り出した文節ごとに変換候補を求め，各文節に対する複数の候補の中から，その文節の周辺にある文節群との共起性が最も高いものを選択することで，その文章の文脈に相応しい変換を導き出すシステムを提案した．漢字の誤変換の検出精度を測定する評価実験を行った結果，変換候補同士の共起性の算定のために参照する上位 Web ページの数 N と，その Web ページ群における参照範囲というパラメータに依存して（$N = 100$ かつ全体参照で）最大 62%という誤字訂正率と，一様に 4%という正字誤訂正率が得られた．

今後は，ある注目している変換候補で Google 検索した際，上位 N 件の Web ページ群を参照して数えた周辺の変換候補の登場回数で算定される共起性が本稿では一方向であるため，注目している変換候補自体に対しても何らかの形で意味合いを持たせ，双方向の共起性として定義することを検討する．また，本稿の提案手法では，1 つの Web ページに変換候補が複数回登場した場合，1 回登場する度に登場回数を加算しており，1 つの Web ページ参照での登場回数カウントが偏り過ぎてしまう危険性もあるため，1 つの Web ページ参照での登場回数カウントの上限を設定するなどによって偏向を防ぎたいと考えている．

文　　献

[1] BABA, 文章校正ツール, http://so-zou.jp/web-app/text/proofreading/ (2015).
[2] 記事作成代行ドットコム, 日本語校正サポート, http://www.kiji-check.com/ (2015).
[3] Microsoft, Word Online, http://office.live.com/start/Word.aspx (2015).
[4] ChaSen, http://chasen-legacy.osdn.jp/ (2015).
[5] Yahoo!, かな漢字変換 API, http://developer.yahoo.co.jp/webapi/jlp/jim/v1/conversion.html (2015).