

平成 27 年度 卒業研究論文

題目 文章校正における共起語を用いた
漢字の誤変換の検出に関する研究

指導教員 服部 峻

提出者 室蘭工業大学 情報電子工学系学科

氏 名 梶谷 貴士

学籍番号 12024035

提出年月日 平成 28 年 2 月 12 日

目次

第 1 章	序論	1
1.1	研究背景	1
1.2	目的	2
第 2 章	関連研究	4
第 3 章	提案手法	5
3.1	システムの構成	5
3.2	入力文の各文節における変換候補の割り出し	5
3.3	各変換候補と周辺の文節との共起性の算定	6
3.4	修正案の作成と表示	7
第 4 章	評価実験	10
4.1	評価方法	10
4.2	各パラメータによる誤字訂正率と正字誤変換率への影響	11
4.2.1	上位 N 件に注目した場合	11
4.2.2	ウェブサイトの参照範囲に注目した場合	11
4.2.3	カウント方法に注目した場合	12
4.2.4	影響度をはかる文節の対象に注目した場合	12
4.3	誤変換検出の結果の詳細	13
4.4	既存の文章校正ツールとの比較	13
第 5 章	考察	15
5.1	誤変換検出失敗の原因分析	15
5.2	正字誤訂正の主な原因	15
5.3	誤変換検出の改善手法	15
5.4	正字誤訂正の防止手法	16
第 6 章	結論	17
	謝辞	18

参考文献 19

付録 A 評価実験に使用した 100 文 20

目次

1.1	決定ボタン押下前の最終イメージ	2
1.2	決定ボタン押下後の最終イメージ	2
3.1	入力文をひらがな文に変換	6
3.2	ひらがな文の各文節から変換候補をリストアップ	7
3.3	Web ページ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	8
3.4	共起性に基づく相応しさの評価値を用いた訂正案の作成	9
4.1	パラメータ N に注目	11
4.2	参照範囲に注目	12
4.3	カウント方法に注目	12
4.4	影響度を測る文節の対象に注目	13

表目次

4.1	誤字訂正率と正字誤訂正率	14
-----	------------------------	----

第1章

序論

この項目では、研究背景と研究目的について述べる。

1.1 研究背景

自分で文章を作成した際、その文章に間違いがないか確かめることがある。完成した文章を自分で一から読んで確認する他、インターネット上に数多く存在している文章校正ツールを使用することもある。しかしながら、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、また、その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」 → 「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」 → 「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」 → 「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。

1.2 目的

本研究では、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ことで精度の高い誤変換検出ができるのではないかと考えた。日々増大して行く Web 上にある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。ある文節に対して共起性の高い語（共起語）を見つけ出し、その語を使用した文を改めてユーザに提示する。最終的には図 1.1 のようにユーザが上の入力欄に、校正したい文章を入力し、決定ボタンを押すと、図 1.2 のように下の欄に訂正後の文章が出力される。もし、ユーザによって入力された文章に誤変換と思われる箇所があれば、システムの考える正しい文章を出力した際、訂正部分が赤文字となって出力される。

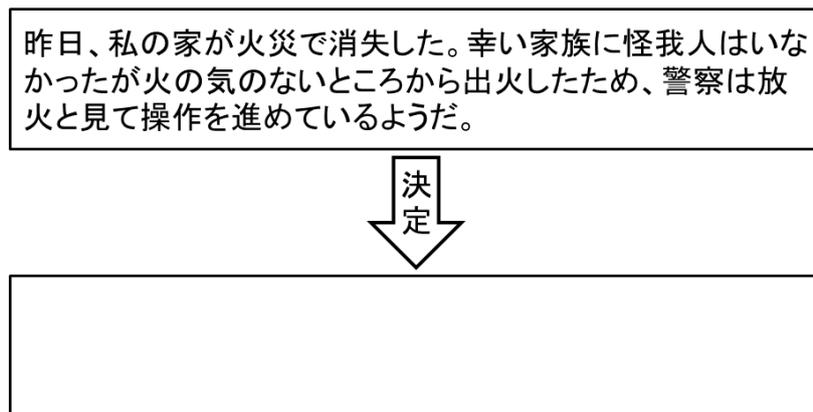


図 1.1 決定ボタン押下前の最終イメージ

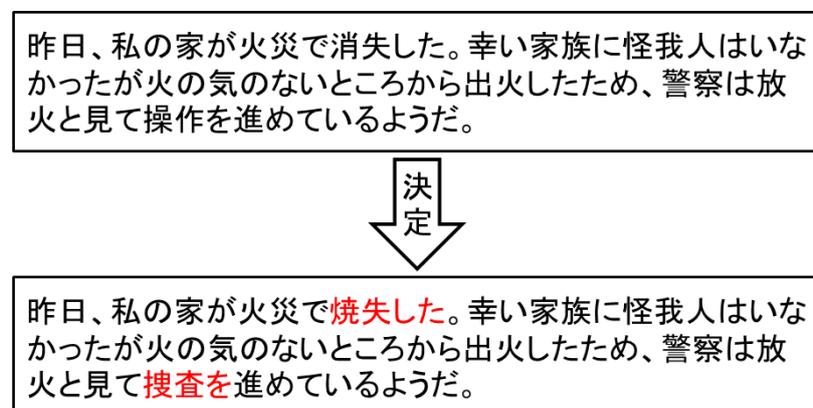


図 1.2 決定ボタン押下後の最終イメージ

また，文章校正の指摘対象の中でも同音異字と同訓異字における誤変換に特化する．主な理由としては，ユーザによって入力された文章に対して，形態素解析することによって文節ごとに分けるためである．例として「おおきなさい」という入力があった場合，他に以下のような変換候補が挙げられる．

- 大きな犀
- お起きなさい
- おお来なさい

しかし，ユーザが文節の壁を越え，上のような間違いを起こすことは考えにくく，かなから漢字に変換した際に違和感を覚えるはずである．よってユーザは文節を越えた誤変換をせず，参考文献 [1] にあるとおり，同音異字や同訓異字における誤変換が一番多いため，そのような誤変換を含む可能性のある文章のみを入力することを前提にする．

第2章

関連研究

この項目では、関連研究について述べる。

山本喜大らの「共起グループを用いたかな漢字変換 [1]」によると、かな漢字変換の誤変換のうち、同音誤りが最も多いと言われている。そこで、直接的あるいは間接的に共起関係にある単語のグループを各分野の専門書から抽出し、共起グループ辞書と呼ばれるものを試作することによって同音語の誤変換の削減の研究をしている。鳥原信一の「テキストの共起単語情報を用いたかな漢字変換 [2]」ではテキストを形態素解析をするのではなく字面から同一の文字種文字列をひとつの単語として切り出し、それらの共起単語情報を納めたテキスト辞書を作成している。これらの研究では、発生しうる同音異義語、同訓異字における誤変換に対して、それぞれ用意した文章から共起語を抽出し、どの変換候補が一番の選択であるかを決定するという点で、本研究と同じである。しかしながら、共起性の高い単語を抽出する際に使用されているテキストの定義が非常に曖昧であるため、どのようなテキストを扱うのかがわからない。また鳥原の研究 [1] では既に完成している本から共起語を抜き出しているが、時代とともに現れる未知語には対応しづらいと考えられる。

第 3 章

提案手法

この項目では，提案手法について述べる．

3.1 システムの構成

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（本研究の実験では文）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．

3.2 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 3.1 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（本研究の実験では文）を形態素解析することにより，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，以下の文節分けの条件

- 文節の中に自立語は 1 つのみ
- 自立語は文節の頭にくる
- 自立語以外はすべて付属語である

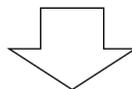
に従い，文節ごとに分かれるように助詞と助動詞以外の，自立語になりうる品詞と判定された単語の直後に空白を入れておくようにする．この空白挿入の処理が必要な理由は，本研究の提案システムで使用している Yahoo!かな漢字変換 API は，ひらがなで入力された文章を文節に分けて，文節ごとの変換候補を求めるものであるが，ひらがなのみで構成された文章では，ユーザからシステムに入力された元の文章の区切り方とは異なる区切り方で区切られてしまう

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 3.2 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、処理して行く。なお、図における「文節 1-1」「文節 2-1」「文節 3-1」「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。この時点で以下のような例外処理を行う。

- 変換候補のうち漢字が一切含まれていないものは除外
- 漢数字ではない数字が含まれているものは除外する
- ひらがなで入力された文節は、ひらがなの文節のみを変換候補とする

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。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。

昨日、私の家が火災で消失した。幸い家族に怪我人はいなかったが火の気のないところから出火したため、警察は放火とみて操作を進めているようだ。



きのう、わたしの いえが かさいで しょうじつした。さいわい かぞくに けがにんは いなかったが ひのけの ないところから しゅっかした ため、けいさつは ほうかと みて そうさを すすすめているようだ

図 3.1 入力文をひらがな文に変換

3.3 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の文節（本研究の実験では前後のみの文節もしくはすべての文節）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 $co(c_{i,j}, c_{i-1,1})$ や、直後の文節に対する 1 番目の変換候補との共起性 $co(c_{i,j}, c_{i+1,1})$ などを算定することで、注目している変換候補 $c_{i,j}$ の

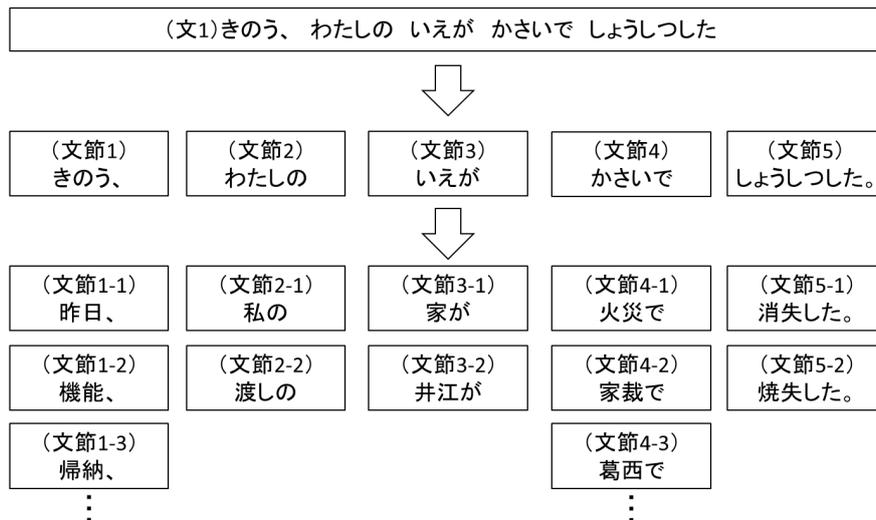


図 3.2 ひらがな文の各文節から変換候補をリストアップ

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いく．具体的には，例えば，図 3.2 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 3.3 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（本研究の実験では $N \in 10, 20, 50, 100$ ）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（本研究の実験では全文参照または注目している変換候補を含む一文参照）にいくつ含まれているかを数えて，前後の文節もしくはすべての文節に対する各変換候補ごとに共起度 $co(c_{3,1}, c_{2,1})$ や $co(c_{3,1}, c_{4,3})$ として登場回数を記録する．但し，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前後の文節に対する各変換候補との共起性を算定する際，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．

3.4 修正案の作成と表示

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 $fitness(c_{i,j})$ を以下の式に基づいて計算し，文節それぞ



図 3.3 Web ページ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

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、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。

- 前後の文節に対する共起性を求める場合

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

- 全文節に対する共起性を求める場合

$$\text{fitness}(c_{i,j}) = \sum_{k=1}^{n_1} co(c_{1,k}, c_{i,j}) \cdots + \sum_{k=1}^{n_i-1} co(c_{i-1,k}, c_{i,j}) + \sum_{k=1}^{n_i+1} co(c_{i+1,k}, c_{i,j}) + \dots$$

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 3.2 の例では、 $n_2 = 2$ 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。最後に、図 3.4 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。

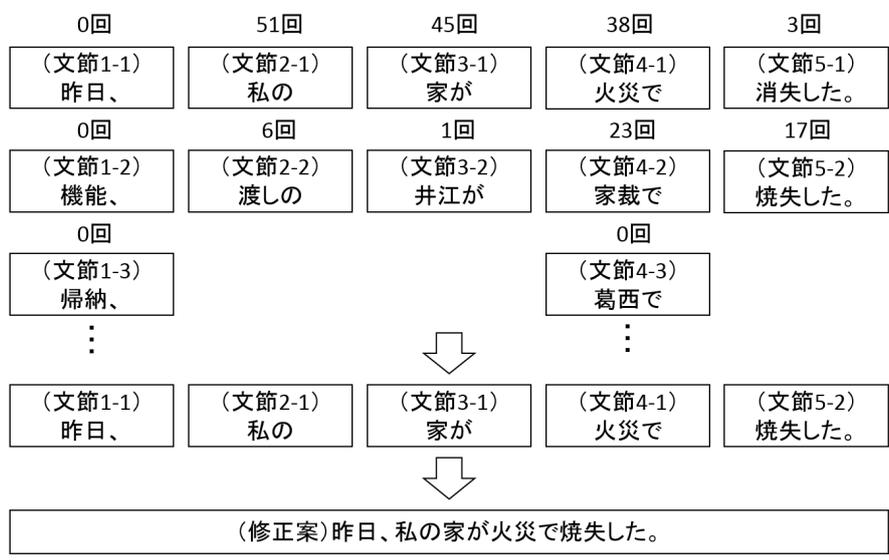


図 3.4 共起性に基づく相応しさの評価値を用いた訂正案の作成

第 4 章

評価実験

この項目では、評価実験について述べる。

4.1 評価方法

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を求めると、評価実験を行うにあたり、システムに対して以下のパラメータとパターンを付与した。

- 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
- 各 Web ページにおける参照範囲に関して
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
- 文節に対する変換候補の総数 1 つの Web ページに複数回変換候補が登場する場合、その総数を記録するパターンと 1 ページ 1 カウントまでで記録するパターン
- 注目している文節が文脈に及ぼす影響の範囲
注目している文節の前後の文節に対して共起性を算定するか、もしくは全文節に対して算定するか

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各パラメータに関して、以上の 4 点について評価実験を行った。従って、計 32 パターンで誤字訂正率と正字誤訂正率を求めた。更に、既存の文章校正ツール Enno[6] に対し同じ 200 文をかけ、精度の比較を行った。

4.2 各パラメータによる誤字訂正率と正字誤変換率への影響

各パラメータ1つずつに注目し、それぞれのパラメータがどのように誤字訂正率や正字誤訂正率に影響を及ぼしているかを見ていく。そのために、注目するパラメータ以外のパラメータを固定し、注目するパラメータにおける実験結果の平均値を求め、その比較を行いグラフ化した。

4.2.1 上位 N 件に注目した場合

Google 検索した際に参照するウェブサイトの上位 N 件について注目した図 4.1 を見てみると N の値を増やすことによって誤字訂正率が大きく上がっていることがわかる。ただし、正字誤訂正率も僅かながら上昇している。

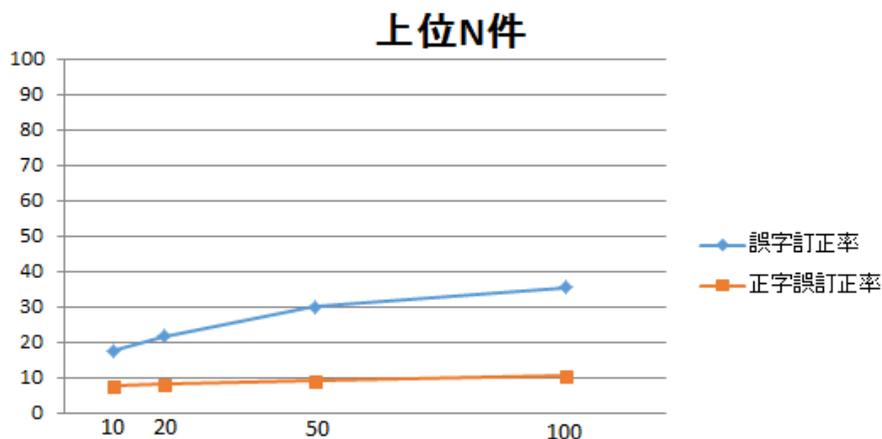


図 4.1 パラメータ N に注目

4.2.2 ウェブサイトの参照範囲に注目した場合

共起性を見つける際の参照範囲について、図 4.2 を見ると一文のみの場合よりもサイト全文参照した場合の方が、より誤変換の検出ができています。しかし上位 N 件の場合と同じように、正字誤訂正率も上昇しています。

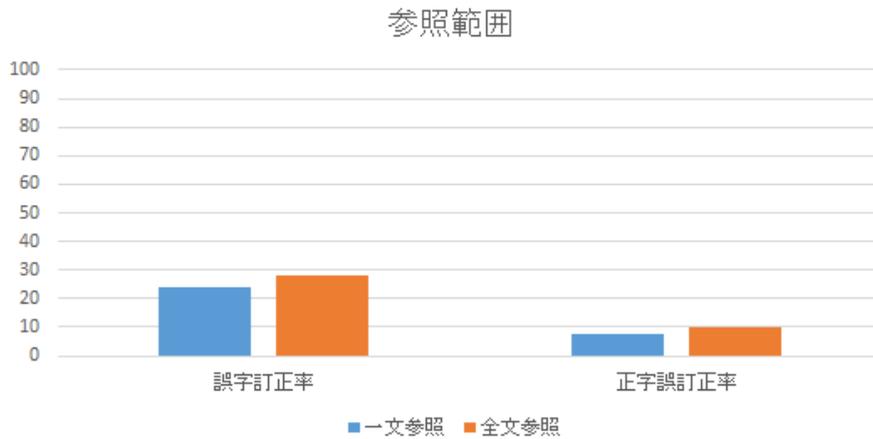


図 4.2 参照範囲に注目

4.2.3 カウント方法に注目した場合

指定した参照範囲に対象の文節の変換候補が見つかった場合、1つのウェブサイトに対して1カウントのみとするか、登場回数をそのまますべて記録するかのパラメータについて図 4.3 を見ると、誤字訂正率、正字誤訂正率ともに変化がないため、よりシステムに負担がかからず、ウェブページによる偏りを防ぐことのできる1カウントが良いと考えられる。

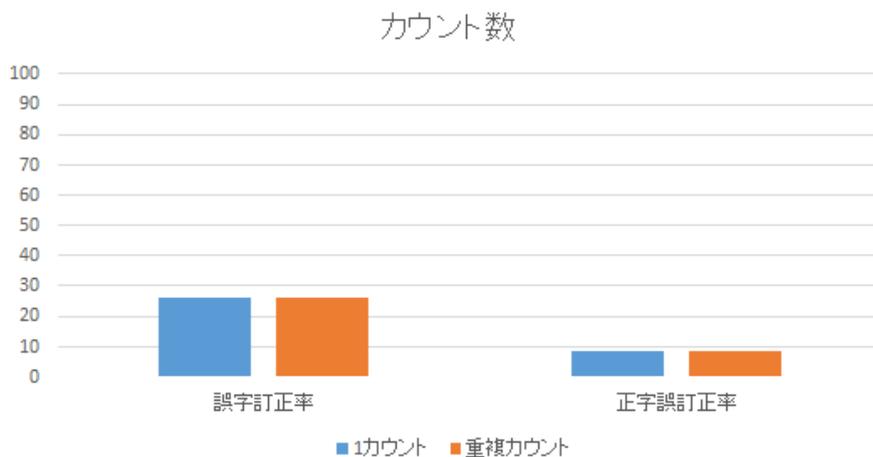


図 4.3 カウント方法に注目

4.2.4 影響度をはかる文節の対象に注目した場合

最後に、影響度を測る文節を前後のみとするか、全文節に対する影響度とするかの場合についての図 4.4 を見ると、前後の文節のみよりも全文節を見る方が若干誤字訂正率が高い。また、正字誤訂正率はどちらも変わらないため、全文節を見る方が良いと考えられる。また、本

実験では短い文が多かったが長い文章になれば単語数も増え、それだけ共起性を導き出すための材料も増えるので、その観点から考えても全文節に対する影響度を確かめる方が良い。

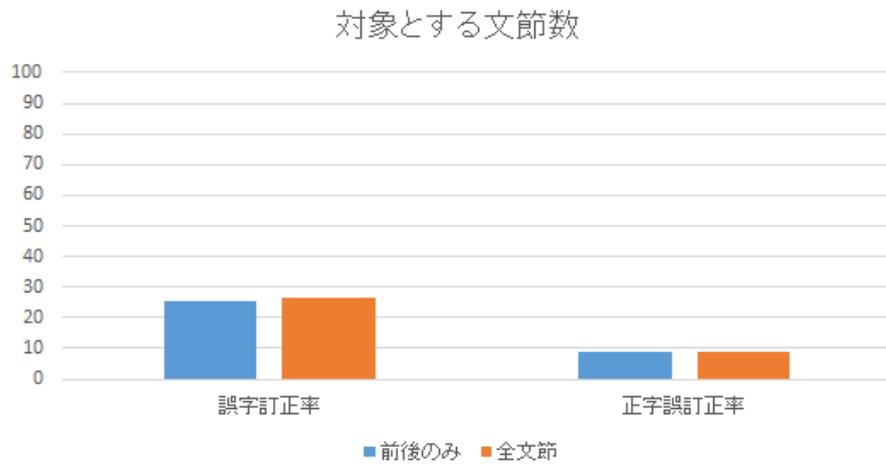


図 4.4 影響度を測る文節の対象に注目

4.3 誤変換検出の結果の詳細

評価実験の結果の詳細は表 4.1 のとおりである。

4.4 既存の文章校正ツールとの比較

本システムとの比較対象として、インターネット上に存在する既存の文章校正ツールの中から漢字の誤変換に特化したツール Enno[6] に今回使用した 200 文をかけてみたところ、誤変換を含む 100 文すべて誤変換なしと判定され、この時点で、本システムがすべての場合で、このツールを上回る結果となった。しかしながら、正字誤訂正も Enno[6] では 0% であったので、その点に関しては劣っていると言える。

表 4.1 誤字訂正率と正字誤訂正率

パラメータ N	参照範囲	記録数	文節数	誤字訂正率	正字誤訂正率
10 件	一文参照	1 カウント	前後のみ	15%	7%
			全文節	16%	7%
		重複	前後のみ	15%	7%
			全文節	16%	7%
	全文参照	1 カウント	前後のみ	19%	8%
			全文節	20%	8%
		重複	前後のみ	19%	8%
			全文節	20%	8%
20 件	一文参照	1 カウント	前後のみ	19%	7%
			全文節	20%	7%
		重複	前後のみ	19%	7%
			全文節	20%	7%
	全文参照	1 カウント	前後のみ	23%	9%
			全文節	24%	9%
		重複	前後のみ	23%	9%
			全文節	24%	9%
50 件	一文参照	1 カウント	前後のみ	27%	8%
			全文節	28%	8%
		重複	前後のみ	27%	8%
			全文節	28%	8%
	全文参照	1 カウント	前後のみ	32%	10%
			全文節	33%	10%
		重複	前後のみ	32%	10%
			全文節	33%	10%
100 件	一文参照	1 カウント	前後のみ	33%	9%
			全文節	34%	9%
		重複	前後のみ	33%	9%
			全文節	34%	9%
	全文参照	1 カウント	前後のみ	37%	12%
			全文節	38%	12%
		重複	前後のみ	37%	12%
			全文節	38%	12%

第 5 章

考察

この項目では、実験を行った結果の考察と、改善手法について述べる。

5.1 誤変換検出失敗の原因分析

誤変換検出の失敗で一番多かった原因は、文節同士の共起性を見つけることができなかったことである。そのため、正しい変換で入力した場合はそのまま正しい変換、間違っただけで入力された場合は間違っただけの変換で出力された。また、Yahoo!かな漢字変換 API により導き出される変換候補の中に正しい変換が存在しない場合もあり、その場合も正しく指摘することができなかった。

5.2 正字誤訂正の主な原因

正字誤訂正してしまった一番の原因は単純に正しい変換よりも誤った変換が多くカウントされ、正しく機能しなかったことである。また、それ以外の原因として、入力された文章を形態素解析しひらがなの文章にする際、ユーザの意図する読み方とは違った読み方で判別されてしまい、その後に導き出した変換候補そのものが変わってしまったことによるものや誤変換検出の失敗と同じように正しい変換候補がかな漢字変換 API に存在していなかったことが挙げられる。

5.3 誤変換検出の改善手法

本システムにおいては、文節同士の共起性を求める際、一方的な共起性しか確かめることができない。文節同士を比較する場合、互いの変換候補すべての組み合わせで共起性を求めておらず、ユーザによって入力された変換との比較であるため、一部の場合で誤変換を検出できなかったと考えられる。例として、「紅顔の少年」という文の一部があった場合、これを以下のように 2 つのパターンで誤変換したとする。

- 「紅顔の少年」 → 「紅顔の正念」、「抗癌の少年」

左の誤変換の場合、「紅顔の」で Google 検索し上位サイトを見ると「しょうねん」の変換候補では一番「少年」が多いため誤変換を指摘できる。しかし、右のような誤変換をした場合、「少年」で Google 検索し上位サイトを見ても「こうがんの」の変換候補である「紅顔の」は出て来ることはなく、誤変換を指摘できずそのままの変換で出力することとなる。つまり、一方の文節の誤りであれば指摘できるがもう片方が間違っている場合は指摘することができない。今後は、双方の共起性を導き出すため、ユーザによって入力された変換と近傍する文節の変換候補に共起性を全く見出すことができなかつた場合、その文節が不適切であると判断し、ユーザによって入力された変換とは別の変換候補で検索し直し、再び近傍する文節との共起性を求めていくといった手法を追加したいと考えている。

ただし、今回は Google 検索を行う際に「Google Custom Search API[6]」を使用したがる、無料で使用する場合、この API には厳しい使用制限があり、1 日で検索できる回数に制限が存在するので、再検索を行った場合にその上限に達してしまう可能性があり、さらに、検索結果も 100 件までしか求めることができないのでパラメータ N の値もこれ以上増やすことができないため、他の検索サイトを利用するなどを検討していきたい。また、変換候補を導き出す際、Yahoo!かな漢字変換 API だけでなく、さらに別のかな漢字変換 API やかな漢字変換辞書を複数用意し、より多くの変換候補を求めることによって正しい変換がそもそも変換候補に出現しないという問題を解決できると考えられるため、導入することを検討している。

5.4 正字誤訂正の防止手法

現状において、ユーザによって入力された文の読み方は形態素解析によるため、完璧に把握することは困難であると考えられ、根本的な解決方法は未だ見つかっていない。しかし、それ以外の原因による正字誤訂正は誤変換検出の改善手法によって、より多くの共起性を抽出することで減少するのではないかと考えられる。

第 6 章

結論

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。そこで本提案では、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した。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。結果として最大 38% の誤字訂正率となり既存の文章校正 Enno[6] を上回る結果となり、ウェブサイトから共起語を導き出し、漢字の誤変換を検出することの有用性が示された。今後は、共起性を発見できず誤変換を指摘できなかった点について、一方的な共起性だけではなく、双方向における共起性を発見し、その情報を利用することによって更なる精度の向上をはかりたいと考えている。

謝辞

本研究に際して、様々なご指導を頂きました服部峻助教を初めとして、服部研究室の皆様に感謝を致します。そして、本研究で用いたフリーの API を提供している二社に感謝致します。

参考文献

- [1] 山本 喜大, 久保田 淳市 : 共起グループを用いたかな漢字変換:情報処理学会, 全国大会講演論文集, pp.189-190 (1992).
- [2] 鳥原 信一 : テキストの共起単語情報を用いたかな漢字変換:情報処理学会, 全国大会講演論文集, pp.225 - 226 (1993).
- [3] Yahoo!日本語形態素解析 API : <http://developer.yahoo.co.jp/webapi/jlp/ma/v1/parse.html>
- [4] Yahoo!かな漢字変換 API : <http://developer.yahoo.co.jp/webapi/jlp/jim/v1/conversion.html>
- [5] Enno : <http://enno.jp/>
- [6] Google Custom Search API : <https://developers.google.com/custom-search/?hl=ja>

付録 A

評価実験に使用した 100 文

評価実験に用いた 100 文を記載する。文中のカタカナ部分に対して、その直後に 2 種類の変換候補が用意されており、左側が正しい変換、右側が誤った変換である。

1. 雪のケッシュョウ（結晶/決勝）はとても小さくて美しかった。
2. 多発している詐欺行為に対して注意をカンキ（喚起/換気）する。
3. 文章ではなく、コウトウ（口頭/高騰）での説明のみであった。
4. 大手企業のサンカ（傘下/酸化）に入ることが決定した。
5. 家宅捜索を行い証拠品をオウシュウ（押収/応酬）する。
6. 完成した絵手紙にラツカン（落款/樂觀）を押す。
7. 基準のカゲン（下限/加減）を下回る数値。
8. 新規参入をソガイ（阻害/疎外）する行為を繰り返す。
9. 一般的にはそのように考えるのがダトウ（妥当/打倒）だ。
10. この絵は有名な画家の絵とコクジ（酷似/告示）している。
11. 落ち着くまでは事態をセイカン（静観/生還）しよう。
12. 彼は両親の愛情にウ（飢/植）えている。
13. 担当者にそのムネ（旨/胸）を伝えるために向かった。
14. 清潔を保つために毎日センジョウ（洗浄/煽情）している。
15. 状況を自分の上司にチュウシン（注進/忠臣）する。
16. 相手と握手とホウヨウ（抱擁/法要）を交わす。
17. 地域のハケン（覇権/派遣）をかけた戦いが始まる。
18. 先日投稿された文章をカイコウ（改稿/開講）する。
19. 残念ながらその提案にはシュコウ（首肯/趣向）できない。
20. 彼はキンセイ（均整/金星）のとれた体格をしている。
21. ユウシ（有志/融資）を募って組織を結成する。
22. 昨年父は長年勤めた会社をテイネン（定年/諦念）退職した。
23. 土地がヒヨク（肥沃/比翼）で作物がよく実る。
24. 手のひらに小さなハンテン（斑点/反転）ができる。

25. 容疑者のシッソウ（失踪/疾走）により捜査は行き詰まった。
26. 売れ過ぎてセイサン（生産/精算）が全く追いつかない。
27. シュヨウ（腫瘍/主要）を取り除く手術を受ける。
28. 相手の要求をイッシュウ（一蹴/一集）することとなった。
29. いきなりショセン（初戦/所詮）を白星で飾る。
30. 私の祖父は高齢であるが、まだまだソウケン（壮健/双肩）である。
31. 日頃の効果的な運動でソウシン（瘦身/送信）する。
32. 今後のドウコウ（動向/瞳孔）に注意するよう言われた。
33. 上司が部下をシッセキ（叱責/失跡）する現場。
34. 台風の影響で川がハンラン（氾濫/反乱）する。
35. 突然の友人のフホウ（訃報/不法）に接する。
36. 農民達は領主に反抗してホウキ（蜂起/放棄）した。
37. 制限の緩和をトクレイ（特例/督励）として認める。
38. 父にイッカツ（一喝/一括）されてすぐに引き下がった。
39. 二つの棒をヘイコウ（平行/平衡）に並べて置きなさい。
40. ホウショク（飽食/宝飾）の時代にも飢餓に苦しむ人がある。
41. 自分勝手な行動で周囲のハンカン（反感/半官）を買う。
42. 利益の一部を客にカンゲン（還元/換言）する。
43. 宝石のカンテイ（鑑定/官邸）を依頼することにした。
44. 部屋の窓を閉めてシャオン（遮音/謝恩）する。
45. 一族のチョウロウ（長老/嘲弄）として尊崇を集める。
46. 彼らは生活のキュウジョウ（窮状/球状）を訴えている。
47. 彼は趣味がコウショウ（高尚/交渉）だ。
48. 円柱のタイセキ（体積/堆積）を求める計算問題。
49. 人材の質にジュウテン（重点/充填）を置く。
50. 寺社の改修のためにジョウザイ（浄財/錠剤）を募る。
51. 必死に子どもがガンカ（玩菓/眼科）を欲しがらる。
52. バイオリンのキョウホン（教本/狂奔）を読むところだ。
53. 彼は未だに大学にセキ（籍/席）を置いている。
54. 経費のセツゲン（節減/雪原）に取り組む。
55. 戦争のせいで物価がトウキ（騰貴/登記）した。
56. 彼の行動はエッケン（越権/謁見）行為に当たる。
57. コウバイ（勾配/購買）の急な坂に気をつける。
58. シュウチ（周知/羞恥）の事実であることを彼女は知らない。
59. その現場にはケツコン（血痕/結婚）が付着していた。
60. ショウガイ（生涯/障害）をかけて達成したい目標ができた。
61. 子供たちのヒーローがカイジュウ（怪獣/懐柔）と戦う。
62. 長年保っていたキンコウ（均衡/近郊）が破られる。

63. 通気性に優れたセンイ（繊維/戦意）で作られた洋服。
64. ひらがなをローマ字にヘンカン（変換/返還）する手間が惜しい。
65. 言葉ではケイヨウ（形容/掲揚）できないほど美しい光景。
66. 周りの意見にカビン（過敏/花瓶）に反応する。
67. 昨日牛肉の輸入がカイキン（解禁/開襟）された。
68. 会社のテイカン（定款/諦観）を作成する。
69. 薬のヨウホウ（用法/養蜂）を正しく守る。
70. 両氏の関係の修復にフシン（腐心/普請）した。
71. 自然の恵みをキョウジュ（享受/教授）する。
72. 三年ぶりに実家にキセイ（帰省/寄生）する。
73. 先生の指摘をシンシ（真摯/紳士）に受け止める。
74. 彼の発言をヨウゴ（擁護/養護）する者はいなかった。
75. 犯人のシモン（指紋/諮問）を採取することに成功した。
76. 学校のコウテイ（校庭/行程）で休み時間にサッカーをする。
77. 週末はどここの工場もカドウ（稼働/可動）していない。
78. 自分のカラ（殻/唐）に閉じこもってばかりいる。
79. ソウチョウ（早朝/総長）から深夜まで営業しているお店。
80. 登山時はセツケイ（雪渓/設計）に注意が必要だ。
81. レーダーで未確認の飛行機をホソク（捕捉/補足）する。
82. 賃金をソキュウ（遡及/訴求）して支払う。
83. 子どもにハシ（箸/橋）の持ち方を教える。
84. 彼とはキュウチ（旧知/窮地）の仲だから安心できる。
85. 自室のキンコ（金庫/禁固）にお金を入れる。
86. 工具が使用されたコンセキ（痕跡/今夕）が残っている。
87. 友人との間にカンゲキ（間隙/感激）が生じた。
88. 道路の夜間コウジ（工事/公示）が行われている。
89. 彼女は有名な女子大出身のサイエン（才媛/再演）である。
90. ジミ（地味/滋味）な服装の男性が多い。
91. 仕事によるヒロウ（疲労/披露）が身体に蓄積している。
92. 予定の飛行機にトウジョウ（搭乗/登場）できなかった。
93. スミ（墨/隅）と筆を用いて文字を書く。
94. 防災対策が選挙の重要なソウテン（争点/掃天）となっている。
95. コンサートホールには大勢のカンキャク（観客/閑却）が詰めかけた。
96. 裁判所にソジョウ（訴状/遡上）を提出する。
97. 無理やり本をジヒ（自費/慈悲）で出版する。
98. 彼が乾杯のオンド（音頭/温度）を取る。
99. 多くのフサイ（負債/夫妻）を抱えるはめになった。
100. 腕の皮膚がエンショウ（炎症/延焼）を起こした。