漏れのない漢字変換誤り検出と誤り可能性によるレベル分け

林 秀治　　山本 和英

長岡技術科学大学

{hayashi, yamamoto}@jnlp.org

1. はじめに

近年、文章を作成するときはワードプロセッサを用いることが一般的となっている。ワードプロセッサで漢字を入力するときは、一度ひらがなで入力してから漢字に変換するので同音異義語への変換誤りが起こりやすい。そのため、変換誤りを対象とした自動検出や自動訂正を行う研究が行われてきたがそのほとんどが高再現率、高精度を目指すものである。実際に自動誤り検出や訂正のシステムを使用する場合すべての誤りを検出、訂正できていなければ、結局ユーザはすべての文章を自分の目で確かめなければならないので負担の軽減にはなりにくい。

そこで、本研究では漢字の変換誤りを検出対象とし、誤りを検出しなかった文を確認する必要がない、漏れのない誤り検出システムの作成を目的とする。

2. 関連研究

以前の研究[1]では、保険文書中の変換誤りを漏れなく検出することを目指した。しかし、この研究では検出対象を内容語としたことで、誤りを含む文を入力したときに、形態素解析の誤りが原因で誤り部分を検出対象として抽出するとができず、検出漏れが生じてしまった。奥ら[2]は、形態素解析による誤りを回避するため機械的に収集した文字の連鎖を手掛かりとし同音異義語の誤りを検出する手法を提案している。しかし、この手法は複合語中の同音異義語を対象としており、再現率も100%は達成できていない。川原ら[3]は形態素解析の誤りを、文字 trigram を使った単語分割法で回避し、コーパスから抽出した単語で辞書を作成し、辞書中に存在しない単語が出現した場合誤りとするという方法で検出を行っている。この手法では、正しい文を誤りとすることはほぼないが再現率は 60~80%ほどとなっている。

3. 提案手法

変換誤りは同じ読みの表記が複数ある場合、つまり同音異義語をもつ場合のみに生じる。そこで、本研究では読みと漢字表記の対の辞書を作成し、それを基に検出を行う。しかし、同音異義語をもつ漢字すべてを誤りとして検出すれば、変換誤りをすべて検出することができるが、ほぼすべての文が誤りとなってしまう。そのため、漢字を含む単語とその前後に出現する語の辞書を作成し、それを基に絞り込みを行う。

3.1 単語分割方法

通常の形態素解析では、誤りを含んだ文を入力すると解析に失敗することがある。解析に失敗すると誤り検出の失敗にも繋がるので今回は初めから形態素解析器を使う方法(分かち書き)と、あらかじめ文中から漢字のみを抜き出し、その漢字に対して形態素解析を行う方法(漢字分割)の 2 種類の方法で分割を行った。形態素解析には MeCab(1)を使用した。形態素解析では分かち書きのみを行い、品詞や読みなどは使用しない。

3.2 『漢字-読み』、『読み-漢字』辞書

同音異義語を検出するために『漢字-読み』辞書と『読み-漢字』を作成する。辞書の作成には IPA 品詞体系辞書(2)を使用した。『漢字-読み』辞書は漢字を含む単語と、その単語の全ての読みの対からなり、中身は例 1 のようになっている。『読み-漢字』辞書は『漢字-読み』辞書で登録された読みと、その読みを持つすべての漢字を含む語の対からなり、内容は例 2 のようになっている。

後,あと,こう,うしろ,のち,ご

例 1：『漢字-読み』辞書の内容

あと,後,跡,痕,蹟,安登,阿戸

例 2：『読み-漢字』辞書の内容

3.3 2gram 辞書の作成

同じ読みで表記が複数あるものでも対象語の前後を確認すれば一通りに絞れる場合がある。例えば、『覧』という漢字は同じ読みの漢字に『乱』という漢字がある。この 1 文字だけでは表記が複数あるため誤りとなってしま

うが『ご覧ください』という使われ方をされていたとき、『乱』という漢字は『ご乱ください』とはまず使われないので絞り込むことができる。

この絞り込みを行うために、漢字を含んだ対象語、その前後に出現した語とその回数の対の辞書を作成する。辞書は正しい文を分割し、漢字を含む語とその前後を抽出することで作成する。単語分割は分かち書き、漢字分割の 2 種類で行い、分割方法毎に別の辞書を作成し、分かち書きでは前後の形態素、漢字分割では前後が漢字を含む場合はその形態素、ひらがなであれば直近の 1 文字のみを登録する。

辞書を作成するための文書は正しい文であるとともに分野に偏りがあってもいけないので『現代日本語書き言葉均衡コーパス』(BCCWJ)[4]から辞書を作成した。

3.4 誤り検出

誤り検出は 1 文単位で、以下の手順で行う。

1. 入力文を分かち書きで単語に分割する
2. 先頭から順に単語を見ていき漢字を含む単語があれば、それを対象語として『漢字-読み』辞書から読みを取得する
3. 取得した全ての読みの全ての表記を『読み-漢字』辞書から取得する
4. 対象語の前後の語が 2gram 辞書に登録されているか調べ、登録されていなければ誤りあり、登録されていてかつ全ての読みで表記が 1 つであれば誤りなし
5. 前後が登録されていて、いずれかの読みで表記が複数の場合、今見ている前後の語が他の表記の前後としても 2gram 辞書に登録されているか調べる。他のいずれの表記でも登録されていなければ誤りなし
6. 誤りなしとなった文について漢字分割で単語に分割し 2.~5.の手順で再度検出を行う

いずれかの手順で、対象語が辞書になければ誤りとなる。検出の例を図 1 に示す。

この例では、対象語が『幹事』のときに手順 4.で後ろの語の『文字』が 2gram 辞書内に存在しないため誤りとなる。入力文が『漢字文字列』であれば、『漢字』の後ろの『文字』は 2gram 辞書内に存在するため手順 5.に進み、他の読みが『かんじ』の表記については 2gram 辞書内に『文字』が存在しないため誤りとならない。

また、以下の 11 語は頻度が高く、絞り込みには向かないと判断したため、前後に出現した場合、例外として 5.の処理を行わなかった。

か,と,に,の,は,へ,も,や,を.が,で

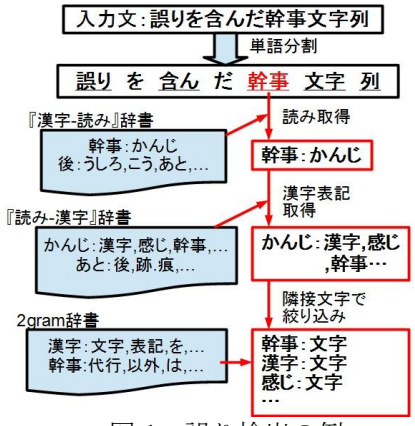

図 1：誤り検出の例

4. 実験

4.1 実験データ

BCCWJ は 13 種類のレジスターからなるので、各レジスターから 1,000 文ずつをランダムに抽出し、この 13,000 文を実験用の誤りのない文とした。今回は漢字変換誤りが検出の対象で誤りであるため、漢字を含む文のみを抽出した。また、一般の文書の評価のため、無条件に各レジスターから各 100 文ずつ、計 1,300 文を抽出した。この 14,300 文を除いた残りをすべて辞書作成に用いた。

誤り文は抽出した誤りのない 13,000 文から以前の研究[1]でテストセットを作る際に用いた方法と同様に、MeCab と IPA の情報を使って単語を同じ読みの別の単語に置換することで作成した。今回は、漢字を含む語のみが検出対象なので、置換は漢字を含む語から漢字を含む語への置換のみで行った。また、誤りは 1 文中に 1 つのみである。置換の結果 327,588 文の誤りを含む文が生成された。この誤りのない 13,000 文と誤りを含む 327,588 文をテストセットとした。

無条件で抽出した 1,300 文については、漢字を含む語という条件はなしに置換を行い、98,519 文の誤り文を作成し、これらをオープンテストセットとした。

4.2 比較手法

比較手法として、同じ読みの表記が複数ある語を含む文を誤りあり、表記が一つのもののみを含む文を誤りなしとするベースラインと機械学習を使った方法を用意した。

機械学習には SVM-Light(2)を使用した。訓練データにはテストセットとオープンテストセットの文を除いた BCCWJ と、テストセットの誤りのない文を、漢字を含む語という条件をなしに置換した誤り文から各 25 万文ずつ抽出したものを使用した。各要素は分かち書きした単語とし、その単語の文中の割合をその素性値として使用した。

4.3 実験および結果

SVM の学習にテストセットの一部を使用しているため、オープンテストセットを入力として提案手法、ベースライン、SVM の 3 つの手法で誤り検出を行った。結果を表 1 に示す。

表 1：オープンテストセットの検出結果

	誤りあり		誤りなし	
	誤りなし	誤りあり	誤りなし	誤りあり
提案手法	54	98,465	26	1,274
ベースライン	47	98,472	22	1,278
SVM	17,447	81,072	1,152	150

この結果を見ると SVM は誤りのない文を誤りなしとするには優れているが、誤りを含む文も多く誤りなしとしてしまうので、誤りを漏れなく検出するという目的には不向きであることがわかる。ベースラインの結果を見ると、ほぼすべての文を誤りありと判定している。このことから、ほぼ全ての文で表記が複数の語が出現していることがわかる。提案手法の結果はベースラインのものと大きな差は見られず、検出漏れもいくつかあることがわかる。これはオープンテストセットでは誤りとなる置換部分が、検出対象とならない漢字を含まない語になることがあるためである。しかし、誤りなしの文もベースラインとほとんど差がなく提案手法が優位であるとは言いがたい。そこで、誤りありとなった文についてさらに絞り込みを行う処理を追加する。

5. レベル分け処理の追加

今までの手法で誤りありとなった文についてさらに処理を行い、誤りなしと判別するのは難しい。そこで、誤りである可能性の高さからレベル分けを行い、高レベルの文では誤りなしの文が少なく、誤りありの文が多くなるようにする。絞り込みには 2gram の情報だけでは不十分であったので、同文中に出現する語の情報を使用する。

5.1 共起辞書の作成

2gram のみでは絞り込みは不十分であったので、同文中に出現する漢字を含む語の辞書を作成する。正しい文を単語に分割し、漢字を含む語すべてを共起したことのある語として登録する。この辞書を作るときの分割方法は分かち書きのみで、漢字分割では作成しない。辞書の作成には、テストセットとオープンテストセットを除いた BCCWJ を使用した。

5.2 レベルの定義

誤りありと判定された文について、誤りの可能性が高いものほどレベルが高くなるようにレベルを設定した。レベルは 0~9 の 9 段階である。定義を以下に示す。

0. 検出対象なし(漢字を含まない文、または全ての漢字を含む語が辞書中になし)
1. 表記が複数で一つに絞れないときに、それぞれ前後の語との組み合わせの頻度を比較し、対象語との組み合わせが一番頻度が高い
2. 表記が複数で前後が 3.4 の例外処理の語、または文の先頭や末尾で存在しない
3. 表記が複数で前後のいずれかの語が 2gram 辞書にあるが 1 つに絞れない
4. 表記は 1 つだが前後いずれかの語が 2gram 辞書にない
5. 表記複数かつ前後いずれかが 2gram 辞書にない

6~9 はレベルが 2~5 のときに、誤りとされた語と同文中の単語の共起の割合を共起辞書で調べ、一定以下だったものがレベルが 4 上がったものになる。例えば、レベル 2 の条件を満たし、誤りとされた語と、同文中の共起辞書に登録されている語の割合が一定以下であればレベルは 4 上がり 6 となる。

図 1 の例では、『かんじ』の表記が複数かつ『文字』が 2gram 辞書にないためレベルは 5 となる。また、同文中の漢字を含む単語『誤り』『含ん』『文字』『列』と共起したことがあるか調べ、例えば『含ん』のみと共起したことがある場合共起の割合は 0.25 となる。もし、共起の割合が 0.25 以下の文のレベルを上げるならば、この文のレベルは 4 上がり 9 となる。

5.3 テストセットでの評価

テストセットを入力とし、共起の割合が 0.5 未満のときにレベルを上げるとし、誤り検出を行った。結果を表 2 に示す。

結果を見るとまず、漢字を含む語への置換では検出の漏れがないことがわかる。レベル分けの結果では誤りありの文はレベル 9 が大半を占め、誤りなしの文でのレベル 9 は 3 割以下であることがわかる。

また、共起の割合が 0.1、0.25 未満と 0 のときにレベルを上げる場合についても同様の評価を行った。結果をグラフにまとめたものを図 2 に示す。

まず、割合が 0、つまり共起した語が一つもない場合のみレベルを上げたときを見ると、誤りのない文を入力したときでも共起した語がない場合が多くあり、誤りのある文で

表 2：レベル分けを行う場合の誤り検出結果

	誤りなし	0	1	2	3	4	5	6	7	8	9
誤りなし	223	25	99	675	1,095	510	382	2,190	4,027	645	3,129
誤りあり	0	31	107	647	3,170	3,223	8,929	11,339	69,246	5,807	225,089

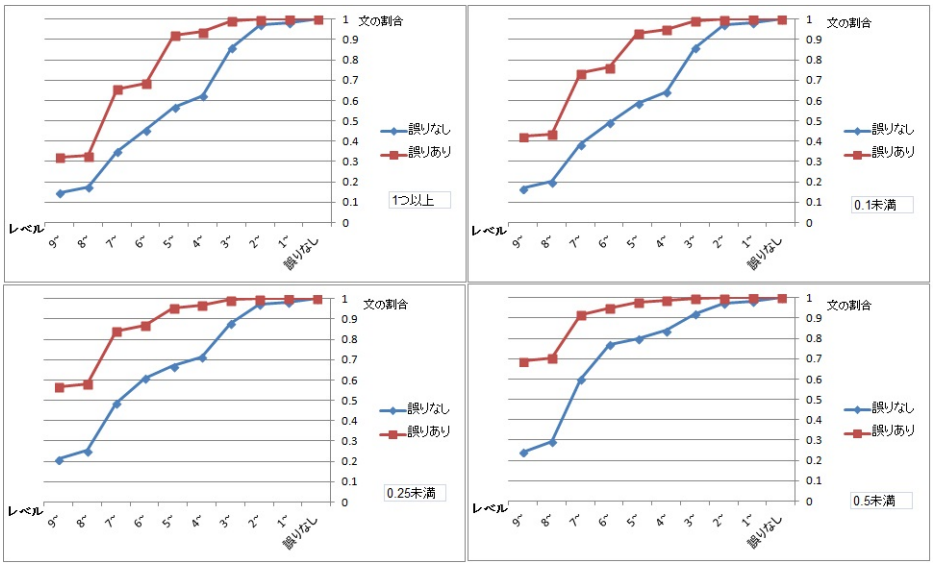

図 2：割合を変化させた場合のレベル毎の文の割合

共起した語がある場合も多くあることがわかる。0.1、0.25 未満を見ると程よくレベル分けがされていて、とくに 0.1 未満の場合ではレベル 4 以上の文を見れば、誤りなしの文が全体の 6 割弱に対し、誤りありの文は全体の 9 割以上を見ることができ、効率よく誤り文を確認できることがわかる。割合が 0.5 未満のときはあまり変化がなく、レベル 6 の段階で多くの誤りなしの文が含まれてしまうことがわかる。割合を 0.75、0.9 未満とした場合でもあまり変化は見られなかった。

6. おわりに

本研究では、実際にユーザが使用することを想定した、漏れのない漢字変換誤り検出システムを作成した。同音異義語が存在する語を誤り、存在しない語を誤りとしない方法をベースとし、検出対象の語の前後の語や同文中の共起したことのある単語の割合を用いて絞り込みを行った。その結果、漢字を含む文については、すべての誤りを検出することができた。また、誤りが検出された文をレベル分けすることにより、全体の 9 割以上の誤りを確認する際に、誤りでない文が全体の 6 割弱しか含まれないような効率よく誤り文を確認できる検出を行えた。

使用した言語資源及びツール

(1) IPA 品詞体系辞書 IPADIC,Ver.2.7.0,
奈良先端科学技術大学院大学松本研究室,
http://sourceforge.jp/projects/ipadic/

(2) 形態素解析器 MeCab, Ver.0.98,
http://mecab.sourceforge.net/

(3) SVM-Light
http://svmlight.joachims.org/

参考文献

[1] 林秀治, 山本和英.保険関連文書を対象とした文章校正支援のための変換誤り検出. 言語処理学会第 20 回年次大会, pp.618-621,2014

[2] 奥雅博, 松岡浩司, "文字連鎖を用いた複合語同音異義語誤りの検出手法とその評価" 言語処理学会論文誌,Vol.4,No.3, pp.83-99,1997

[3] 川原一真, 山本幹雄,"コーパスから抽出された辞書を用いた表記誤り検出法" 情報処理学会第 54 回(平成 9 年前期)全国大会, pp.21-22,1997

[4] 山崎誠[編『書き言葉コーパス]-設計と構築-』講座日本語コーパス 2,朝倉書店, 2014(ISBN978-4-254-51602-9 C3381).