6ZB-07

キー入力の癖の学習に基づく入力間違い推定法の検討

西村 希槻† 寺澤 卓也†
†東京工科大学 メディア学部 メディア学科

1 はじめに

近年，PC を使用して，数千字の文書を書く機会は少なくない．特に PC 操作が不得意であると誤入力を起こす可能性が考えられる．そのため文書校正を行う必要性は高い．

ワードプロセッサには，辞書に基づいて単語の校正を行う機能がある．しかし，登録されていない単語は校正できない．そこで本研究では，辞書に依存せず，ユーザごとのキー入力の癖から間違えやすいキーを推測して校正する方法を考案した．

2 提案手法

2.1 概要

本手法では，まず，ユーザにサンプル文書を入力してもらい，まとまった量のキー入力をキーロガーで取得する．次に作成したプログラムで，取得したログデータから誤入力に関連したデータを抽出し，機械学習が行える形に変換する．そして機械学習で癖のパターンを学習させ，確率の計算を行う．最後に作成した学習データに基づき，校正を行う文書に対して，確率によって間違えやすいと判断された箇所を中心に，ハイライト表示を行い，校正の支援を行う．

2.2 ユーザの入力の取得

機械学習を行うためには，まとまった量のキー入力を修正動作も含めて取得する必要性がある．キー入力を取得するため，キーロガーを使用する．これにより機械学習を行うために十分なログデータを蓄積させることができる．

2.3 ログデータ中の必要なデータの抽出と変換

次にログデータ中に不要なデータがあることによって学習結果に影響を及ぼさないよう，誤入力のみの抽出を行う．ログデータから誤入力部分を抽出するには，BackSpace，もしくは Delete キーで訂正された部分を手がかりとする．同じ文字を何度も訂正していればその文字の入力をユーザが苦手としている可能性があると判断できる．図 1 の場合，m を打とうとして，a と誤って入力し m に訂正したことが考えられるため，a と[BackSpace]を抽出する．

K
A
[BackSpace]
M
A

図 1：打ち間違いの例

また， pen と入力するつもりが，誤って oen と入力し，3 文字消して訂正した場合，e と n は削除して入力し直しており，誤入力した文字と訂正した文字が同じとなる．間違いであるのは p と打とうとして o と誤って入力した箇所であるため，この場合の e と n の修正は誤入力ではないと見なし，除外する．これにより，誤入力の箇所のみを抽出できる．

次に，使用するライブラリは，学習させるデータを読み込ませる際，二次元配列の形に変換する必要があるため，Python のプログラムを作成し，抽出した文字を二次元配列に変換する．

2.4 機械学習

本研究では機械学習環境として，Python のライブラリであり，トピックモデルと呼ばれる文章中の単語の出現確率を推定できる gensim[1]を使用する．gensim は他の機械学習ライブラリと比較すると図 2 の位置づけだと考えられる．

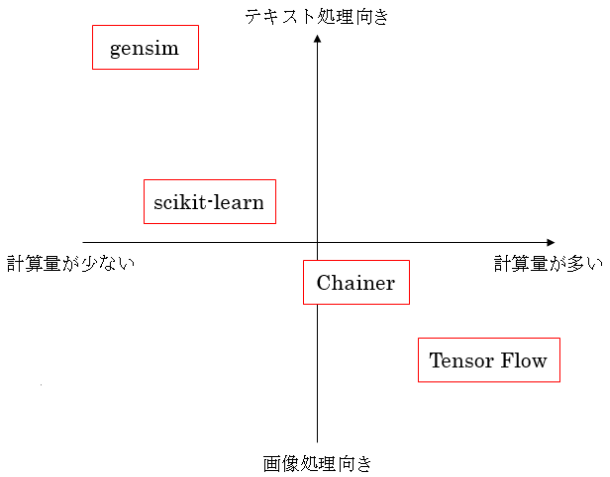

図 2:機械学習ライブラリの比較

A presumption method to find typos based on machine-learning of typing habits
Kizuki NISHIMURA† and Takuya TERASAWA†
†School of Media Science, Tokyo University of Technology

本研究では，gensim のクラスの一つで，独立した単語同士が，設定回数以上出現したら，1 つの単語として結合し扱うことができる Phrases[2]を使用する．これを，誤入力のパターンが一定回数出現した場合，誤入力の癖として認識させるために用いる．これにより，1 回しか間違えなかったといった偶発的な誤入力を除去する．また，間違いのパターンの出現回数をカウントし，以下の数式で誤入力の確率を算出する．

$$\text{あるキーを誤入力する確率} = \frac{\text{そのキーの誤入力の回数}}{\text{そのキーの入力回数}}$$

2.5 文章構成の手法

校正自体は人間が行う．校正対象の文書のうち，以上の処理によって判明したチェックすべき箇所を示すために，複数の単語を一度にハイライトできるテキストエディタを使用する．その際，確率を参照し，確率毎に 3 レベルにハイライト表示を色分けして行うことで，特に注意深く見るべき箇所を把握しやすくした．

3 検証と実験結果

3.1 1500 文字での検証

ユーザのキー入力の癖を取得するため，まずある程度の量のテキストを入力してもらいキーロガーで記録した．ユーザの使用環境を考慮し，キーボードは普段使用しているものを使用するなどを行い，癖以外で起こる誤入力を防いだ．また，キーロガーは日本語を記録できないため，ログデータを取得するためには，pykakasi[3]によって日本語をローマ字化したものを入力する形にした．

まず 1500 文字で行った結果，文字数が少ないため，偶発的に起こった間違いを除去することができなかった．しかし，検証のためにユーザに文書校正を行わず作成させたテキストを解析すると，あらかじめ人間が調査し発見した誤入力のうち，75%の誤入力を本研究の手法で検出することができた．

3.2 10000 文字での検証

次に行った実験は，前回の結果を踏まえ 10000 文字の入力で行った．検証を行うために入力をしたユーザから，PC を使用して実際に過去に講義で作成したレポートの文書を提供してもらい，学習結果を反映させた．目視で調査することによって検出した結果との比較から，10000 回文字入力で 3 回以上間違えた文字を誤入力のパターンとすることで，偶発的に起きる間違いを取り除けると判断した．

3.3 ハイライト表示の検証

まず，偶発的な誤入力のみ取り除き，2.5 の手法でハイライト表示を行った結果，ほとんどの文字がハイライトされてしまった．そのため，ある文字について，誤入力の回数に応じて設定したしきい値を，誤入力する確率の値が超えていれば，その文字をハイライトすることにした．今回は，10 回以上間違えた文字であれば誤入力確率 2%以上の文字をハイライト，10 回以下であれば 3%以上の文字をハイライトするようにした．その結果，図 3 のようにハイライトする箇所を減らすことができた．

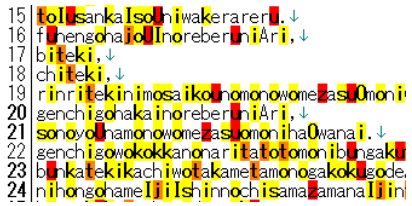

図 3：ハイライトする文字を限定した具体例

4 結論

本研究の手法，すなわち，キーロガーによりキー入力を取得し，それに含まれる誤入力を元にした機械学習によって，校正対象文書中の誤入力をある程度検出できることが分かった．また，ハイライト表示の手法では，誤入力回数に応じて決めたしきい値を，算出した確率の値が超えていれば，その文字に対し，ハイライト表示を行うことで適切に校正支援できるめどをつけられた．

5 今後の課題

本手法によって校正する文章中の入力ミス発見の精度をどこまで上げることができるのかは明らかになっていない．そのため，文字をどの程度解析することが適切なのかを明確にする必要がある．また，解析したユーザの他の文書に対してハイライト表示を行った場合，同じ精度で検出することが可能なのかを明確化できなかったため，今後，検証する必要がある．

参考文献

[1] Radim Řehůřek,“genism” https://radimrehurek.com/gensim/models/phrases.html 最終閲覧日 2016/12/13

[2] Radim Řehůřek,“Phrases“ https://radimrehurek.com/gensim/models/phrases.html 最終閲覧日 2016/12/13

[3] miurahr,“pykakasi” https://github.com/miurahr/pykakasi 最終閲覧日 2016/12/19