

素朴な疑問

英文スペリングコレクタで正しい単語候補の 選び方はどうなっているの？†

平 川 秀 樹††

いくつかの英文スペリングコレクタ*を比較してみるとすぐに分かりますが、同じスペリング誤りに対して、システムによって、種々の異なった訂正単語候補が得られます。また、示される候補の数も、比較的少ないものから、かなり多いものまで様々です。これはスペリング訂正の方式、規則、辞書、ノウハウなどがそれぞれ異なっているからです。残念ながら、商用システムについては、処理方式やノウハウは、公表されておらず、それらの振舞いを説明できるだけの情報を提供することはできません。ここでは、英文スペリングコレクタの方式をいくつか紹介し、実際のシステムに取り入れられると推定されるノウハウ的知識を紹介します。

英文スペリングコレクタは（1）テキストからスペリング誤りの可能性のある文字列（誤り候補）を取り出すスペリングチェックと、（2）誤り候補に対する訂正単語候補となる文字列（訂正候補）を生成するスペリング訂正の2つの作業を行います。

スペリングチェックに関しては、ほぼすべてのシステムで、ノンワードエラー検出（nonword error detection）という方式が採用されています。これは、そのシステムの英単語辞書に登録されていない文字列を、誤った単語の候補として検出する方式です。この方式では、本来誤りである文字列でも、それがたまたま辞書に存在していると、誤りとして検出できないという限界があります。たとえば、“form”を“from”と誤った場合がこれにあたります。この種の誤りは、リアルワードエラー（real word error）と呼ばれ、現在

* How Do English Spelling Correctors Work? by Hideki HIRAKAWA (Communication and Information Systems Research Laboratory, Toshiba Research and Development Center).

† (株)東芝研究開発センター情報・通信システム研究所

‡ 通常、誤り単語の指摘を行うソフトをスペリングチェック、訂正まで行うソフトをスペリングコレクタと区別して呼びます。

の研究課題となっています。ノンワードエラー検出は、基本的にテキスト中の単語文字列を辞書と照合することで実現できます。辞書との照合には、効率的な辞書検索法（ハッシュ法、バイナリサーチ法、トライ構造など）や、辞書をビット列に変換しておき単語の有無のみを高速判定する手法などが採用されています。

スペリング訂正は、誤り候補に「類似した」辞書中の単語を訂正候補として取り出してくる処理であり、いくつかの方式が開発されています。ここでは、2つの代表的手法について触れますか、興味のある方は文献1)のサーベイを参照ください。

(1) Minimum edit distance 法

文字列間の距離を定義して、誤り候補と辞書中の単語の距離を計算し、距離の近いものを訂正候補とします。

一文字削除（“stringe” → “string”）、一文字追加（“strng” → “string”）、一文字の置き換え（“streng” → “string”）、および隣接文字の反転（“strign” → “string”）をそれぞれ1つの訂正操作とします。この操作の適用により、任意の文字列間の変換ができます*。この変換に必要な最小の操作数をその文字列間の距離（Damerau-Levenshtein 距離）と定義します。たとえば，“strang”に対して、“string”，“strange”は距離が1，“stringy”，“straying”は、距離が2となり、人間の感覚にかなり合っているといえるでしょう。この距離は、ダイナミックプログラミングにより、文字列長の2乗のオーダで計算でき、また、距離が1（=誤り数が1）を仮定すれば、文字列長のオーダで可否を判定できます。一般的に、Minimum edit distance 法では、辞書のすべての単語との距離を計算をしなければなりませんが、これは、計算量的に現実的ではなく、一般に単語

* 4つの訂正操作が任意の文字列変換に必須であるというわけではありませんが、これは、後述するようにタイプミスの主要要因に関連付けて設定されています。

候補の絞り込みなどが必要となります。

DEC-10 のスペリングコレクタ他いくつかのシステムでは、Reverse minimum edit distance 法を採用しています。これは誤り候補から 1 回の訂正操作を適用して生成可能な文字列を生成し、それらの文字列のうち辞書に載っている単語を訂正候補とする方式です。訂正操作を 1 回とする理由は、1 回の操作でもスペリング誤りの約 80% が回復可能であり、また、2 回以上の操作では生成される文字列の数が極端に増えてしまうためです。実際、10 文字の文字列に対して、生成可能な文字列数は、1 回の訂正操作では、555 ですが、2 回の訂正操作では、約 30 万となります。この方式では、生成される候補は誤り数が 1 のもののみですので、訂正候補の数は、かなり少なくなります。

(2) Similarity key 法

Similarity key 法は特定の規則に従って文字列をキー (key) にマップしたときに、同一あるいは類似したキーを持つ文字列を類似文字列として判定する方法です。この方法は、辞書の単語に対するキーを事前に計算し、インデックスを作成しておくことが可能ですので高速に類似単語を検索することができます。

SPEEDCOP²⁾ というシステムでは、スケルトンキー (skelton key) というキーが使用されています。スケルトンキーは、もとの単語のアルファベットを、(a) 最初の 1 文字、(b) 子音を重複なく出現順、に並べて構成します。たとえば、“chemical” という語は、“c”+“hml”+“eia”，すなわち、“chmleia” というキーを持ちます。辞書中のすべての単語のスケルトンキーを事前に計算し、アルファベット順にソートしてテーブル化しておきます。類似単語の検索は、このテーブルを誤り候補のスケルトンキーで検索し、同一ないしは、最も類似するキーの前後一定範囲に存在するキーを持つ単語を候補として取り出すことにより行われます。

スケルトンキーは、単語の最初に表れる文字と子音に敏感で、これらの抜けの検出に弱いという弱点があります。このため、SPEEDCOP では、省略キー (omission key) というキーも別途使用しています。省略キーは、子音の省略を考慮したキーで、人間の直感とはかなりかけ離れた単語

を類似単語として判定します。たとえば、“microelectronics” と “circumstantial” は、非常に類似した省略キーを持ちます。こうしたキーの使用は、人間では考えられない訂正候補が示される一要因となる可能性があります*。

この他、訂正候補の推定には、単語や文字列の並びに関する統計データに基づく確率・統計モデルによる手法、誤りを修正する経験的知識を用いて訂正候補を生成する規則ベースの手法などが開発されています¹⁾。

一般に、スペリング訂正の精度や応答速度を上げるための最適化を行う際には、人間のキータイプ誤りに関する知識が役立ちます。実データに基づく調査も行われ、たとえば次のような経験的知識が知られています。

(1) エラーの約 80% は、一文字の抜け・追加・打ち間違いおよび隣接文字の反転である。一文字の抜けが 30% と最も多い。

(2) 最初の文字を誤るというエラーは少ない。

(3) キーボード上で隣接したキーの打ち間違いや隣接キーの同時押しが多い。

(4) 子音の重ね誤りは、起りやすい。

たとえば、(3)を使用すれば、隣接キーの打ち間違いを正す訂正候補を優先して提示できます。また、キーボードの配列などを考慮して、打ち誤りにくい文字のペアを訂正の対象から除外して処理の簡略化を図り、応答速度を上げることも可能です。ただし、このような処理は、生成する候補を限定しますので、ユーザが期待する単語候補が示されなくなる要因にもなります。この他、システムによって、単語の出現頻度により処理を変更したり、発音の類似性を考慮したりする (“sorser” に対して “saucer” を提示) など、種々のノウハウが組み込まれています。

参考文献

- 1) Kukich, K.: Techniques for Automatically Correcting Words in Text, ACM Computing Surveys, Vol. 24, No. 4 (1992).
- 2) Pollock, J. and Zamora, A.: Automatic Spelling Correction in Scientific and Scholarly Text, Communications of the ACM, Vol. 27, No. 4 (1984).

(平成 7 年 8 月 14 日受付)

* これは、一般論であり、SPEEDCOP がこうした訂正候補を出力するというわけではありません。



平川 秀樹（正会員）

1956年生。1980年京都大学大
学院工学研究科電気工学専攻修
士課程修了。同年東京芝浦電氣
(株) (現、(株)東芝) 入社。以

来、機械翻訳、談和解析などの自然言語処理システムの研究開発に従事。1982~1985年(財)新世代コンピュータ技術開発機構研究員。1993~1994年(株)日本電子化辞書研究所第五研究室室長。1994~1995年MIT メディアラボ研究員。現在(株)東芝研究開発センター情報・通信システム研究所主任研究員。人工知能学会、言語処理学会、ACL各会員。

「素朴な疑問」をお寄せください

学会誌編集委員会

学会誌編集委員会では、会員・読者との交流をはかるため、このコーナーを開設いたしました。皆様から日頃疑問に感じている情報処理、計算機科学、計算機工学に関する素朴な質問を募集いたします。学会事務局あてFax、または電子メールで質問をお寄せください。

Fax.(03)5484-3534
e-mail:editj@ipsj.or.jp